Centrum voor Wiskunde en Informatica

REPORT*RAPPORT*

PNA

Probability, Networks and Algorithms

Probability, Networks and Algorithms

How Mobility Impacts the Flow-Level Performance of Wireless Data Networks

Thomas Bonald, Sem Borst, Alexandre Proutiere

REPORT PNA-R0401 JANUARY 30, 2004

CWI is the National Research Institute for Mathematics and Computer Science. It is sponsored by the Netherlands Organization for Scientific Research (NWO).
CWI is a founding member of ERCIM, the European Research Consortium for Informatics and Mathematics.

CWI's research has a theme-oriented structure and is grouped into four clusters. Listed below are the names of the clusters and in parentheses their acronyms.

Probability, Networks and Algorithms (PNA)

Software Engineering (SEN)

Modelling, Analysis and Simulation (MAS)

Information Systems (INS)

P.O. Box 94079, 1090 GB Amsterdam (NL)
Kruislaan 413, 1098 SJ Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

ISSN 1386-3711

How Mobility Impacts the Flow-Level Performance of Wireless Data Networks

ABSTRACT
The potential for exploiting rate variations to improve the performance of wireless data networks by *opportunistic* scheduling has been extensively studied at the packet level. In the present paper, we examine how slower, mobility-induced rate variations impact the performance at the flow level, accounting for the dynamic number of users sharing the transmission resource. We identify two limit regimes, termed *fluid* regime and *quasi-stationary* regime, where the rate variations occur on an infinitely fast and an infinitely slow time scale, respectively. Using stochastic comparison techniques, we show that these limit regimes provide simple, *insensitive* performance bounds that only depend on easily calculated load factors. Additionally, we prove that for a broad class of Markov-type fading processes, the performance varies monotonically with the time scale of the rate variations. The results are illustrated through numerical experiments, showing that the fluid and quasi-stationary bounds are remarkably sharp in certain typical cases.

2000 Mathematics Subject Classification: 60K25, 68M20
Keywords and Phrases: flow-level performance, fluid regime, insensitivity, mobility, performance bounds, processor sharing, quasi-stationary regime, rate variations, response times, scheduling, slow fading, throughputs
Note: Work carried out in part under the project PNA2.2 `Wireless Networks'.

How Mobility Impacts the Flow-Level Performance of Wireless Data Networks

Thomas Bonald[†], Sem Borst[⋆,‡], Alexandre Proutière[†]

[†]France Telecom R&D
38-40 rue du Général Leclerc, 92794 Issy-les-Moulineaux, France

[⋆]CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

[‡]Bell Laboratories, Lucent Technologies
P.O. Box 636, Murray Hill, NJ 07974-0636, USA

Abstract

The potential for exploiting rate variations to improve the performance of wireless data networks by *opportunistic* scheduling has been extensively studied at the packet level. In the present paper, we examine how slower, mobility-induced rate variations impact the performance at the flow level, accounting for the dynamic number of users sharing the transmission resource. We identify two limit regimes, termed *fluid* regime and *quasi-stationary* regime, where the rate variations occur on an infinitely fast and an infinitely slow time scale, respectively. Using stochastic comparison techniques, we show that these limit regimes provide simple, *insensitive* performance bounds that only depend on easily calculated load factors. Additionally, we prove that for a broad class of Markov-type fading processes, the performance varies monotonically with the time scale of the rate variations. The results are illustrated through numerical experiments, showing that the fluid and quasi-stationary bounds are remarkably sharp in certain typical cases.

1 Introduction

Next-generation wireless networks are expected to support a wide variety of high-speed data applications, in addition to conventional voice services and current low-bandwidth data services such as short messaging. The integration of these heterogeneous applications on a common transmission infrastructure raises similar challenges as in wireline integrated networks. In wireless environments, these issues are further exacerbated by interference

problems, intrinsically limited bandwidth, and highly variable and unpredictable propagation characteristics. Specifically, the channel quality may vary widely among spatially distributed users due to distance-related attenuation. In addition, the channel conditions for a given user may vary dramatically over time because of fading effects.

Fading is an extremely complex physical phenomenon caused by the interaction between the propagation environment and user mobility. It emerges in diverse forms and typically spans a wide range of time scales. Fast fading arises because of multi-path propagation effects, and as the word suggests, occurs at a relatively high pace. Slow fading manifests itself at a more macroscopic level as a result of distance-related attenuation and scattering due to obstacles and terrain conditions, and evolves over a longer time scale.

Wireless circuit-switched voice networks rely on power control mechanisms for adjusting the transmit power to combat fading and maintain a fixed transmission rate. Various data applications on the other hand, such as file transfers and Web browsing sessions, are less sensitive to packet-level delays, and do not have a stringent rate requirement. Such elastic applications are well-suited for rate control algorithms which dynamically adapt the transmission rate over time so as to match the fluctuations in channel quality. The resulting variations in the transmission rates in fact open up the possibility of scheduling data transmissions to the various users when their channel conditions are relatively favorable. While fading is considered to have a predominantly adverse impact for voice connections, it thus provides the opportunity to achieve throughput gains for elastic data transfers.

The performance gains from *opportunistic* scheduling rely on the rates varying sufficiently slowly to be tracked with reasonable accuracy, but relatively fast compared to the delay tolerance of the users. High-frequency fading causes estimation and prediction problems, diminishing the scope for scheduling. Slow variations cannot be harnessed, or only at the expense of compromising the delay allowance of the users. For example, typical values of the time constant in the Proportional Fair algorithm for the CDMA 1xEV-DO system [7, 12, 17] range from 10 to 1000 slots of 1.67 ms. This ensures that starvation effects cannot persist for excessive periods, but it also implies that slower variations are not exploited. In practice, relatively low-mobility scenarios tend to provide the greatest potential for scheduling gains.

While the performance of opportunistic scheduling algorithms has been thoroughly explored at the packet level [1, 3, 10, 20, 24, 27], the impact of fading on flow-level performance has received remarkably little attention so far. In [9], it was shown that when the fading is relatively fast compared to the flow dynamics, the system may in certain cases be modeled as a Processor-Sharing type queue with a state-dependent service rate that accounts for the scheduling gains. This model provides explicit formulas for the distribution of the number of active flows and the mean transfer delay. In particular, the performance is *insensitive*, in the sense that these measures only depend on the statistical characteristics of the system through an easily computed 'load' factor. The notion of 'cell capacity', critical

for dimensioning purposes, can then be defined independently of the detailed properties of the system [8].

In the present paper, we focus on the impact of mobility-induced fading that evolves on a slower time scale and manifests itself in the form of rate variations at the flow level. Due to these slower rate variations, the insensitivity property is lost, and the performance depends in some complicated fashion on the detailed rate statistics and traffic characteristics of the system, rendering an exact analysis impractical. Considering these complexities, we compare the performance of the system with that in two limit regimes, termed *fluid* regime and *quasi-stationary* regime, where the rate variations have the same instantaneous statistics, but occur on an infinitely fast and an infinitely slow time scale, respectively. Using stochastic comparison techniques, we show that the fluid and quasi-stationary regimes yield optimistic and conservative performance estimates, respectively. These estimates are particularly useful, since the performance in the limit regimes *is* insensitive, and only depends on appropriately defined load factors, thus providing simple bounds that render the detailed statistical characteristics of the system largely irrelevant. Numerical experiments indicate that these bounds are surprisingly tight in certain typical cases.

The above ordering results show that infinitely fast rate fluctuations yield the best performance, while infinitely slow variations produce the worst performance. It is tempting to conjecture that the performance improves monotonically as the fading process is sped up. We demonstrate that this is indeed the case for a broad class of Markov-type fading processes.

It is worth observing that the above ordering results relate to a change in the time scale of the rate variations for *given* instantaneous rate statistics. As mentioned above, the actual transmission rates may be reduced at higher fading frequencies because of estimation and prediction problems, so for a given system, a change in the time scale will also affect the marginal rate distribution to some extent.

At a qualitative level, the finding that mobility-related rate variations improve the performance ties in with the generic principle described earlier with respect to opportunistic scheduling. It further resonates with the observation in [14] that mobility increases the capacity of ad hoc wireless networks. In the present context, however, the performance improvement does not rely on opportunistic scheduling, but also occurs for example in case of channel-oblivious round-robin scheduling. Instead, informally stated, it arises from the fact that flow-level performance measures behave as convex functions of the rate processes.

The remainder of the paper is organized as follows. In Section 2 we present a detailed model description. In Section 3 we introduce the fluid and quasi-stationary regimes mentioned above. We establish a necessary and sufficient stability condition in Section 4. In Section 5 it is proved that the fluid and quasi-stationary regimes provide stochastic bounds for the performance of the actual system. For Markov-type fading processes, we demonstrate

in Section 6 that the performance varies monotonically with the time scale of the rate fluctuations. In Section 7 we discuss some numerical experiments performed to illustrate the analytical results. In Section 8 we make some concluding remarks.

2 Model description

We consider a single base station serving a dynamic number of flows, each involving a finite data transfer. The service rate of each flow depends on the channel quality and the number of competing flows. Packet-level dynamics are encapsulated in the way flows share the transmission resource, as explained below. Each flow is characterized by its size (in bits) and its 'feasible' transmission rate that is (slowly) time-varying.

Specifically, we consider an arbitrary number K of flow classes, each class corresponding to given flow size and rate statistics. Class-k flows arrive as a Poisson process of rate λ_k. Denote by F_{ki} the size of the i-th arriving class-k flow, and by $R_{ki}(t)$ its feasible rate at time t, corresponding to the actual service rate of the flow if it were the only one in the system. (For notational convenience, we define $R_{ki}(t)$ for all values of t. Note however that the i-th class-k flow may not have arrived yet or may already be completed at time t, in which case $R_{ki}(t)$ is of no significance.) We assume that F_{ki} and $R_{ki}(t)$, $i = 1, 2, \ldots$, are i.i.d. copies of a random variable F_k and a stationary and ergodic process $R_k(t)$, respectively. The process $R_k(t)$ is assumed to be bounded and right-continuous with left-hand limits.

Denote by $C_k := \mathrm{E}\left[R_k(0)\right]$ the time-average feasible rate of a class-k flow. Define $\rho_k := \lambda_k \mathrm{E}\left[F_k\right]/C_k$ as the traffic load associated with class k and denote by $\rho := \sum_{k=1}^{K} \rho_k$ the total traffic load. It is not directly clear what the right concept of 'load' is in view of the time-varying transmission rates. In particular, the load as defined above does not coincide with the fraction of time that the base station is active. However, the results in Section 4 will show that the above-defined notion does provide a correct measure of load from a stability perspective.

Assuming that the packet-level scheduling results in fair sharing at the flow level, the actual service rate of the i-th arriving class-k flow, if present at time t, is:

$$R_{ki}(t)G(n)/n, \tag{1}$$

where n denotes the total number of flows present at time t. The function $G(n)$ accounts for possible throughput gains from channel-aware scheduling. In particular, the function $G(n)$ with $G(1) = 1$ is increasing in n and tends to some finite limit value G^* for $n \to \infty$, while the ratio $G(n)/n$ is decreasing in n.

Remark 1 *Fair sharing trivially occurs in case of static round-robin scheduling for example, corresponding to $G(n) \equiv 1$, but it may also naturally arise in case of channel-*

aware scheduling. Specifically, in case $R_{ki}(t) \equiv C_k$, the model reduces to that considered in [9] for the flow-level performance of a weight-based scheduling strategy which assigns weights $w_k = 1/C_k$ to class-k flows. In case the flows have statistically identical normalized rate variations $Y_1, Y_2, \ldots$ at the packet level, it may then be shown that $G(n) = \mathrm{E}[\max\{Y_1, \ldots, Y_n\}]$. As may further be deduced from [2, 9, 15, 19, 25], the Proportional Fair algorithm for the CDMA 1xEV-DO system would approximately behave like the above weight-based strategy, provided the exponential smoothing window is sufficiently large. In case the feasible transmission rate $R_{ki}(t)$ is (slowly) time-varying, similar arguments suggest that a weight-based strategy which assigns a dynamic weight $w_{ki}(t) = 1/R_{ki}(t)$ to the i-th class-k flow, results in the actual service rate (1) at the flow level.

Remark 2 *The comparison results to be derived in Sections 5 and 6 in fact remain valid under the even milder assumption that the service rate of the i-th class-k flow is:*

$$R_{ki}(t)H_k(n_1, \ldots, n_K), \tag{2}$$

where n_k denotes the number of active class-k flows and the function $H_k(\cdot)$ is decreasing in each of the n_k's. The above service rate function may be used to model the flow-level performance of a broad class of discriminatory scheduling algorithms. Unfortunately however, when the function $H_k(\cdot)$ is not of the form $G(n)/n$ with $n = \sum_{k=1}^{K} n_k$ as in (1), the fluid and quasi-stationary regimes described below prove extremely difficult to analyze.

3 Definition of fluid and quasi-stationary regimes

The flow-level model defined by (1) corresponds to a Processor-Sharing type queue where the service rates of the various users are modulated by independent stochastic processes. Considering the complexity of the system, we introduce two limit regimes, termed *fluid* regime and *quasi-stationary* regime, where the rate processes evolve on an infinitely fast and an infinitely slow time scale, respectively. Formally, let us consider a family of systems, parameterized by $s \in (0, \infty)$, where the generic rate process for class-k flows is $R_k^{(s)}(t) \equiv R_k(st)$. Thus the parameter s represents the 'speed' of the rate process. Or equivalently, the value $1/s$ models the time scale of the rate process. In case $R_k(t)$ is a Markov process, the process $R_k^{(s)}(t)$ may be obtained by scaling the transition rates with s.

When the parameter s grows large, the rate process approximately averages out over the time scale of the flow dynamics. In the limit for $s \to \infty$, the variations completely vanish, and the rate process reduces to a constant, giving rise to the 'fluid' regime with $R_k^{\mathrm{fl}}(t) \equiv R_k^{(\infty)}(t) = C_k$. On the other hand, as the value of s becomes small, the fading process remains roughly constant over the time scale of the flow dynamics. In the limit for $s \to 0$, the changes completely disappear, and the rate process freezes in some initial state, yielding

the 'quasi-stationary' regime with $R_k^{\rm qs}(t) \equiv R_k^{(0)}(t) = R_k(0)$, where $R_k(0)$ has the stationary marginal distribution of the process $R_k(t)$.

Accordingly, define the traffic loads associated with class k in the fluid and quasi-stationary regimes as $\rho_k^{\rm fl} := \lambda_k \mathrm{E}\left[F_k\right]/C_k$ and $\rho_k^{\rm qs} := \lambda_k \mathrm{E}\left[F_k/R_k(0)\right] = \lambda_k \mathrm{E}\left[F_k\right]/C_k^{\rm qs}$, where $C_k^{\rm qs} := \mathrm{E}\left[1/R_k(0)\right]^{-1}$. Note that these load factors depend on the rate statistics only through the arithmetic and harmonic means, respectively. By Jensen's inequality, we have $\rho_k^{\rm fl} \leq \rho_k^{\rm qs}$. Denote by $\rho^{\rm fl} := \sum_{k=1}^{K} \rho_k^{\rm fl}$ and $\rho^{\rm qs} := \sum_k \rho_k^{\rm qs}$ the total traffic loads in the fluid and quasi-stationary regimes, respectively. Recall that $\rho \equiv \rho^{\rm fl}$.

As mentioned earlier, the fluid and quasi-stationary regimes are particularly relevant, because their performance may be explicitly evaluated. Based on the results of [9, 18], it can be shown that a necessary and sufficient condition for stability of the fluid (respectively, quasi-stationary) regime is $\rho^{\rm fl} < G^*$ (respectively, $\rho^{\rm qs} < G^*$). When the system is stable, the stationary distributions $\pi^{\rm fl}$ and $\pi^{\rm qs}$ of the numbers $(n_1, \ldots, n_K)$ of active flows of the various classes in the fluid and quasi-stationary regimes depend on the traffic and rate statistics through the load factors $\rho_k^{\rm fl}$ and $\rho_k^{\rm qs}$ only:

$$\pi^{\rm fl}(n_1, \ldots, n_K) = \pi^{\rm fl}(0) \frac{n}{\phi(n)} \prod_{k=1}^{K} \frac{(\rho_k^{\rm fl})^{n_k}}{n_k!},$$

$$\pi^{\rm qs}(n_1, \ldots, n_K) = \pi^{\rm qs}(0) \frac{n}{\phi(n)} \prod_{k=1}^{K} \frac{(\rho_k^{\rm qs})^{n_k}}{n_k!},$$

where $n = \sum_{k=1}^{K} n_k$, $\phi(n) := \prod_{i=1}^{n} G(i)$, and $\pi^{\rm fl}(0)$ and $\pi^{\rm qs}(0)$ are determined by the respective normalizing conditions. By Little's law, we obtain the mean response time of class-k flows:

$$\mathrm{E}\left[T_k\right] = \frac{\mathrm{E}\left[n_k\right]}{\lambda_k}.$$

Alternatively, the performance may be naturally measured in terms of the *flow throughput*:

$$\gamma_k := \frac{\mathrm{E}\left[F_k\right]}{\mathrm{E}\left[T_k\right]} = \frac{\rho_k C_k}{\mathrm{E}\left[n_k\right]}.$$

When $G(n) \equiv 1$, we obtain, for the fluid and quasi-stationary regimes, respectively:

$$\gamma_k^{\rm fl} = C_k(1-\rho), \qquad \gamma_k^{\rm qs} = C_k^{\rm qs}(1-\rho^{\rm qs}). \tag{3}$$

4 Stability condition

The system described in Section 2 is said to be stable if, starting from any initial state, it converges to a finite stationary regime. It follows from the stochastic bounds to be derived in Section 5 that the condition $\rho^{\rm fl} < G^*$ is necessary for stability, while the condition

$\rho^{\rm qs} < G^*$ is sufficient. Note that when the number of active flows tends to infinity, each flow stays in the system for a long time, so that the rate process tends to average out over the flow duration, i.e., the system behaves as in the fluid regime. Thus one would expect the condition $\rho \equiv \rho^{\rm fl} < G^*$ to be both necessary and sufficient for stability. The next theorem, proved in Appendix A, confirms that this is indeed the case.

Theorem 1 *If $\rho < G^*$, then the system is stable.*

Remark 3 *It is worth emphasizing that the assumption of fair sharing is crucial for the above stability condition to hold. A weaker stability condition applies in case the scheduling strategy takes advantage of the slow rate variations as well. Observe however that exploiting slow rate variations may cause severe starvation effects and (unacceptably) long flow-level delays. Conversely, one could imagine a perverse non-idling scheduling strategy for which a stronger stability condition might emerge.*

5 Comparison with fluid and quasi-stationary regimes

We now compare the performance of the system with that in the fluid and quasi-stationary regimes, using the notion of stochastic ordering (see for instance [22]).

Definition 1 *(st and icx orderings) Let X and Y be two r.v.'s on $\mathbb{R}^n$. Write $X \leq_{st} Y$ (respectively $X \leq_{icx} Y$) if and only if $\mathrm{E}[f(X)] \leq \mathrm{E}[f(Y)]$ for all increasing (respectively increasing and convex) functions $f : \mathbb{R}^n \to \mathbb{R}$ for which the previous expectations exist.*

Note that these orderings are particularly relevant, since $X \leq_{st} Y$ allows the comparison of the distributions of X and Y, i.e., $\Pr[X \leq x] \geq \Pr[Y \leq x]$ for all x, while $X \leq_{icx} Y$ implies that all moments of X are smaller than the corresponding moments of Y.
Assume that the system is empty at time 0 and denote by $N_k(t)$ the number of active class-k flows at time t. For $i = 1, \ldots, N_k(t)$, let $F_{ki}(t)$ be the remaining size of the i-th active class-k at time t. We define the total workload at time t as:

$$W(t) := \sum_{k=1}^{K} \sum_{i=1}^{N_k(t)} \frac{F_{ki}(t)}{C_k}.$$

Theorem 2 below, proved in Appendix B, states that the performance improves (respectively, deteriorates) in terms of the number of active flows, the workload and the response time T of an arbitrary flow, when the rate processes of some flows satisfying Assumption 1 below are replaced by the corresponding fluid (respectively, quasi-stationary) versions as described in Section 3.

Assumption 1 *The cumulative distribution function (c.d.f.) $P(\cdot) = \Pr[F \leq \cdot]$ associated with the random flow size F is concave.*

Note that Assumption 1 is satisfied by a broad class of distributions, including exponential, hyper-exponential or Weibull. In particular, it is possible to represent the highly variable flow size distribution of typical data networks.

Theorem 2 *We have, for all $k = 1, \ldots, K$,*

$$W^{\mathrm{fl}}(t) \leq_{icx} W(t) \leq_{icx} W^{\mathrm{qs}}(t), \tag{4}$$

$$N_k^{\mathrm{fl}}(t) \leq_{st} N_k(t) \leq_{st} N_k^{\mathrm{qs}}(t), \tag{5}$$

$$T^{\mathrm{fl}} \leq_{st} T \leq_{st} T^{\mathrm{qs}}, \tag{6}$$

where the superscript $^{\mathrm{fl}}$ (respectively, $^{\mathrm{qs}}$) refers to the system where the rate processes of some set of flows satisfying Assumption 1, are replaced by the corresponding fluid (respectively, quasi-stationary) processes.

The above comparison results are also valid when the system is in equilibrium. Denote by $W(\infty)$, $N_k(\infty)$ and $T_k(\infty)$ the workload, the number of active class-k flows and the response time of class-k flows in steady state, respectively. We deduce the inequalities (8) and (9) in the next corollary from Theorem 2 and the stability of the *st*-order by limits [22]. The inequality (7) results from (4) and a classical monotonicity property of the Loynes' construction as explained in [4, page 281].

Corollary 1 *Let $\mathcal{L} \subseteq \{\infty, \ldots, \mathcal{K}\}$ be an arbitrary subset of classes that satisfy Assumption 1. We have, for all $k = 1, \ldots, K$:*

$$W^{\mathrm{fl}}(\infty) \leq_{icx} W(\infty) \leq_{icx} W^{\mathrm{qs}}(\infty), \tag{7}$$

$$N_k^{\mathrm{fl}}(\infty) \leq_{st} N_k(\infty) \leq_{st} N_k^{\mathrm{qs}}(\infty), \tag{8}$$

$$T_k^{\mathrm{fl}}(\infty) \leq_{st} T_k(\infty) \leq_{st} T_k^{\mathrm{qs}}(\infty), \tag{9}$$

where the superscript $^{\mathrm{fl}}$ (respectively, $^{\mathrm{qs}}$) refers to the system where the rate processes of the flows of the classes in $\mathcal{L}$ are replaced by the corresponding fluid (respectively, quasi-stationary) processes.

6 Impact of the time scale of the rate variations

We now investigate how the performance varies with the time scale of the rate variations. In order to do so, we suppose that the processes $R_k(t)$ for some users are replaced by processes $R_k^{(s)}(t) \equiv R_k(st)$ for some constant $s > 1$. The constant s may be interpreted as an acceleration factor. Although one might conjecture that the performance improves when the rate process is sped up, this result does not hold in certain specific cases [23]. However, the monotonicity property can be established when the rate process satisfies the following assumption.

Assumption 2 *The rate process is a homogeneous stationary Markov process. The transition kernels Q and Q_r of the Markov process and of the corresponding time-reversed Markov process are $\leq_{st}$-monotone. Recall that Q is $\leq_{st}$-monotone if and only if, for all increasing functions $f(\cdot)$, the function $x \mapsto \int f(t)Q(x, dt)$ is also increasing [22].*

Assumption 2 is satisfied by a broad class of processes, including birth-death processes and Markov processes with a discrete state space and a generator $Q = (q_{ij})$ such that q_{ij} does not depend on i [6].
The next theorem, proved in Appendix C, states that the performance improves when the rate processes of some set of flows satisfying Assumptions 1 and 2 are accelerated.

Theorem 3 *We have, for all $s > 1$ and all $k = 1, \ldots, K$:*

$$W^{(s)}(t) \leq_{icx} W(t), \tag{10}$$

$$N_k^{(s)}(t) \leq_{st} N_k(t), \tag{11}$$

$$T^{(s)} \leq_{st} T, \tag{12}$$

where the superscript $^{(s)}$ refers to the system where the rate processes of some set of flows satisfying Assumptions 1 and 2, are sped up by a factor s.

The next corollary presents the counterpart of Corollary 1.

Corollary 2 *Let $\mathcal{L} \subseteq \{\infty, \ldots, \mathcal{K}\}$ be an arbitrary subset of classes that satisfy Assumptions 1 and 2. We have, for all $s > 1$ and all $k = 1, \ldots, K$:*

$$W^{(s)}(\infty) \leq_{icx} W(\infty), \tag{13}$$

$$N_k^{(s)}(\infty) \leq_{st} N_k(\infty), \tag{14}$$

$$T_k^{(s)}(\infty) \leq_{st} T_k(\infty), \tag{15}$$

where the superscript $^{(s)}$ refers to the system where the rate processes of the flows of the classes in $\mathcal{L}$ are sped up by a factor s.

7 Numerical experiments

We now present some numerical experiments conducted to illustrate the analytical results. The feasible rate of a user is a complex function depending on both fast and slow fading. For the sake of simplicity in simulations, we ignore fast fading. Slow fading may be viewed as the result of two different phenomena: shadowing and variations in path loss. Thus, we assume that the feasible rate $R(t)$ of a user behaves as:

$$R(t) \propto G_s(t) \times \Gamma(t),$$

Ring j	Rate c_j (Kbit/s)	Radius r_j ($\alpha = 4$)	SINR (dB)
0	2457.6	1	9.5
1	1843.2	1.07	7.2
2	1228.8	1.19	3.0
3	921.6	1.28	1.3
4	614.4	1.41	-1.0
5	307.2	1.68	-4.0
6	204.8	1.86	-5.7
7	153.6	2.00	-6.5
8	102.6	2.21	-8.5
9	76.8	2.37	-9.5
10	38.4	2.82	-12.5

Table 1: Rates, ring radius and SINR for constant shadowing

where $G_s(t)$ and $\Gamma(t)$ denote the shadowing component (of unit mean) and the path loss, respectively. Shadowing typically arises when the variation in the distance to the base station is relatively small, depending on the propagation environment. Empirical studies suggest that shadowing has a log-normal distribution, with standard deviation between 5 and 12 dB. Path loss usually varies over larger distances. Here we assume that $\Gamma(t)$ is proportional to $r(t)^{-\alpha}$, where $r(t)$ denotes the distance to the base station at time t and α is the path loss exponent.
Based on these observations, we consider the following two mobility scenarios:

- *Low* mobility, where the typical distance covered by a user during the flow transfer is relatively limited. Variations in the feasible rate are then mainly due to shadowing and result from fluctuations of 5 to 12 dB in the received signal.

- *High* mobility, where a user may move across the entire cell during the flow transfer. In this case, fluctuations in the feasible rate are mainly due to variations in path loss.

In the CDMA 1xEV-DO system, 11 feasible rates are defined, $c_0 > c_1 > \ldots > c_{10}$, with corresponding target signal to interference-plus-noise ratios (SINR). In case of a constant shadowing component $G_s(t) \equiv 1$, these rates correspond to a set of concentric rings of radius $r_0 < r_1 < \ldots < r_{10}$, such that when $r(t) \in (r_{j-1}, r_j)$, the feasible rate is c_j (with the convention $r_{-1} \equiv 0$) [8]. We give the ring radius (normalized so that $r_0 \equiv 1$) corresponding to a path loss exponent $\alpha = 4$ in Table 1.
In the following, we consider a circular cell of external radius $R = r_L$ corresponding to $L+1$

rings. Flows arrive uniformly in the cell according to a Poisson process of intensity λ. The probability p_j that a new flow starts its service in ring j is proportional to the surface of this ring, i.e., $p_j = (r_j^2 - r_{j-1}^2)/R^2$. The flow throughputs in the limit regimes are given by (3). Simulation results are obtained for exponentially distributed flow sizes of unit mean and Markov rate processes with values in $\{c_0, c_1, \ldots, c_L\}$. We make the natural assumption that the rate process can only jump between adjacent states, so that for each class, the Markov rate process is a birth-death process. Note that Assumptions 1 and 2 are satisfied.

7.1 Low mobility

In the low-mobility scenario, the feasible rate of a user typically evolves in a set of 2 to 5 consecutive rates, roughly corresponding to SINR variations of 5 to 12 dB (see Table 1). Rather than fitting a log-normal distribution, we simply assume that the feasible rate of each user takes a fixed number of values, a, and that all transitions rates of the corresponding birth-death process are equal. We performed simulations not reported here to verify that the performance depends on shadowing mainly through its amplitude and not on its precise distribution. We consider a cell of radius $R = 1.86$ (thus $L = 6$), and evaluate the performance in the following two cases:

- Shadowing with low amplitude ($a = 3$). There are 5 user classes. Class-k users are located in ring k, $k = 1, \ldots, 5$, and their feasible rates are c_{k-1}, c_k, c_{k+1} with corresponding marginal probabilities $\frac{1}{4}$, $\frac{1}{2}$, $\frac{1}{4}$.
- Shadowing with high amplitude ($a = 5$). There are 3 user classes. Class-k users are located in ring k, $k = 2, \ldots, 4$, and their feasible rates are c_{k-2}, c_{k-1}, c_k, c_{k+1}, c_{k+2} with corresponding marginal probabilities $\frac{1}{8}$, $\frac{1}{4}$, $\frac{1}{4}$, $\frac{1}{4}$, $\frac{1}{8}$.

Figure 1 presents the throughput of flows of classes 1 and 5 as a function of cell load in case of shadowing with low amplitude ($a = 3$) and with different values of the speed s. Figure 2 shows the throughput of flows of classes 1 and 3 in case of shadowing with high amplitude ($a = 5$).

As expected in view of Corollaries 1 and 2, the fluid and quasi-stationary regimes provide optimistic and conservative estimates of the throughput, respectively, and speeding up the rate processes improves the performance. Further observe that the limit regimes only differ significantly in case of shadowing with high amplitude.

7.2 High mobility

We now consider a high-mobility scenario where the variations in path loss cannot be neglected. We assume that a fraction β of the users move across the entire cell while the

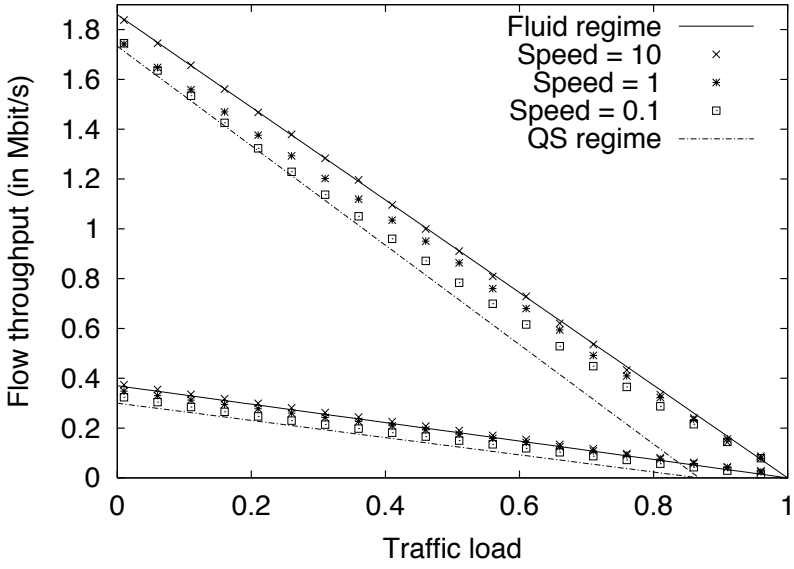

Figure 1: Throughput of flows of class 1 (upper curves) and class 5 (lower curves) in case of shadowing with low amplitude ($a = 3$).

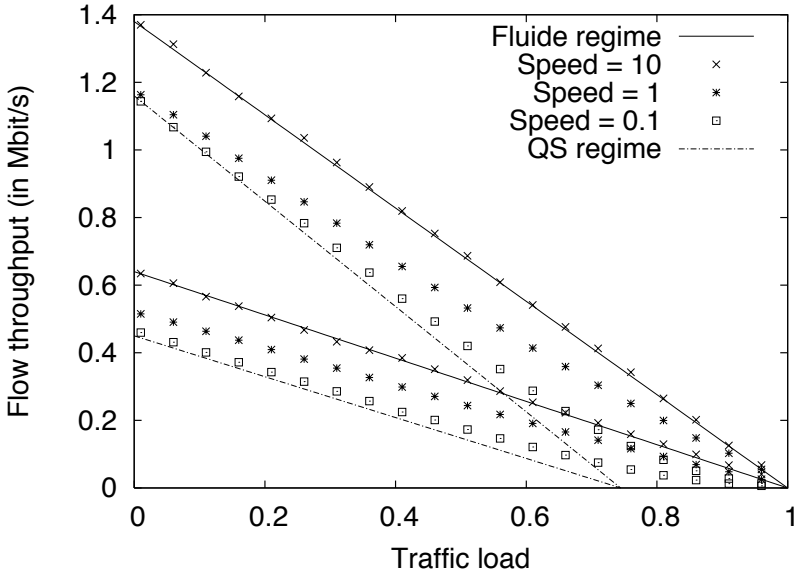

Figure 2: Throughput of flows of class 1 (upper curves) and class 3 (lower curves) in case of shadowing with high amplitude ($a = 5$).

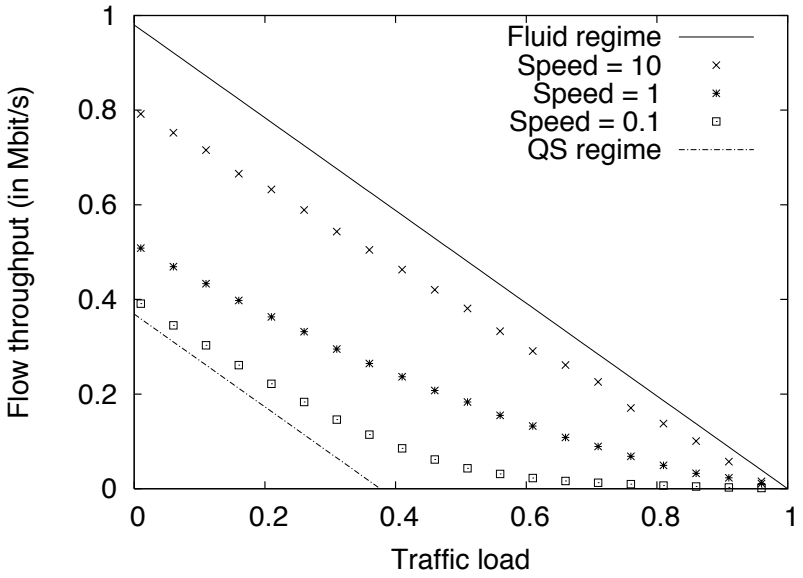

Figure 3: Flow throughput as a function of traffic load when all users move in a cell of radius $R = 2$.

others are static. We do not account for shadowing, i.e., $G_s \equiv 1$. There are $K = L + 2$ user classes:

- Class-k users, for $k = 0, \ldots, L$, are static in ring k, i.e., $R_k(t) \equiv C_k = c_k$ for all t. The load associated with class k is $\rho_k = (1 - \beta)\lambda p_k / C_k$.
- Class-$(L + 1)$ users move in rings $0, \ldots, L$ according to a birth-death process with marginal distribution $p_0, \ldots, p_L$, corresponding to isotropic motion in the cell, so that $C_{L+1} = \sum_{k=1}^{K} p_k c_k$. The load associated with class $L + 1$ is $\rho_{L+1} = \beta\lambda / C_{L+1}$.

Figure 3 gives, for a cell of radius $R = 2$ (thus $L = 7$) where all users move ($\beta = 1$), the flow throughput as a function of total traffic load for different values of the speed s. The impact of speed on the flow throughput for fixed load $\rho = 0.5$ is shown in Figure 4. Figure 5 is the analogue of Figure 3 for a cell of radius $R = 1.19$ (thus $L = 2$). Note that for large variations in the feasible rate (case $R = 2$), performance is highly sensitive to speed, whereas for limited variations (case $R = 1.19$), the fluid and quasi-stationary bounds are very close, indicating that performance is nearly insensitive.
Figure 6 gives the flow throughput of static users in ring 0 and of moving users, for equal fractions of static and moving users ($\beta = 0.5$).
The numerical results suggest that the performance is sensitive to the speed of the fading process only when variations in the feasible rates of those users representing a significant

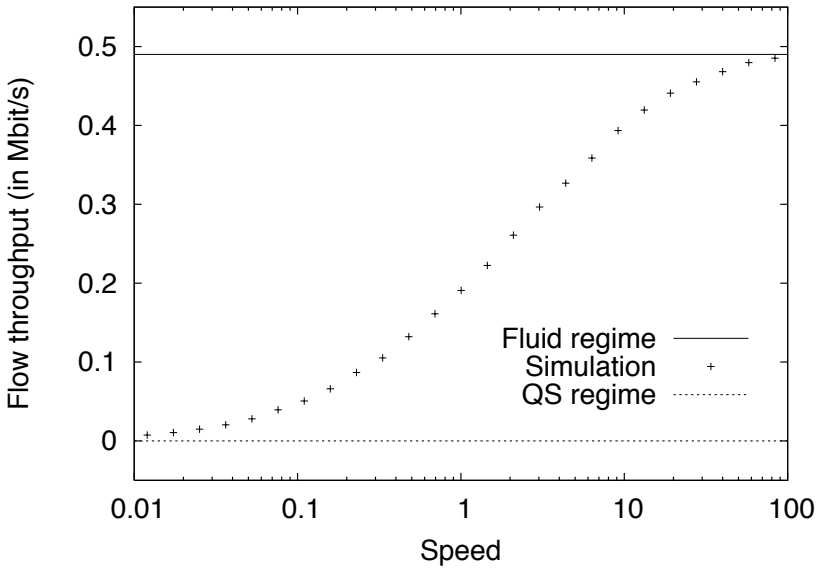

Figure 4: Flow throughput as a function of speed for a fixed traffic load $\rho = 0.5$ when all users move in a cell of radius $R = 2$.

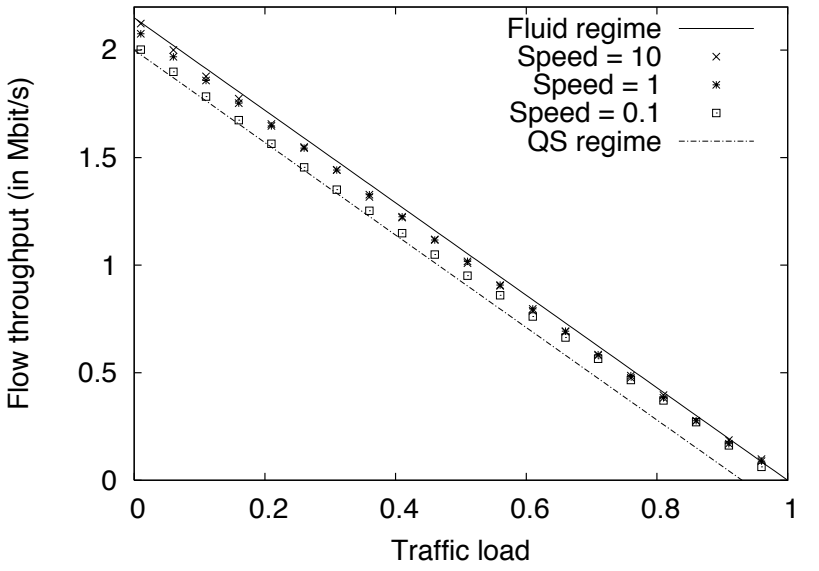

Figure 5: Flows throughput when all users move in a cell of radius $R = 1.19$.

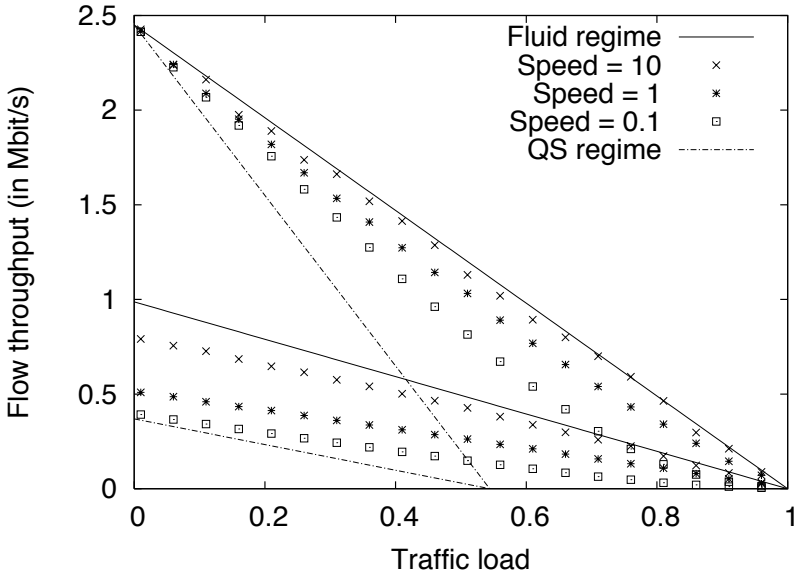

Figure 6: Flow throughput of static users in ring 0 (upper curves) and of moving users (lower curves) in a cell of radius $R = 2$.

part of the total traffic load are of high amplitude. When the amplitude is low, the performance is mostly insensitive, i.e., mainly depending on the traffic and fading statistics through he traffic load of each class only. In this case, the quasi-stationary regime provides an accurate conservative estimate of the flow-level performance.

8 Conclusion

We have examined how slow, mobility-induced rate variations impact the flow-level performance of a wireless data system. We have identified two limit regimes, termed fluid regime and quasi-stationary regime, where the rate variations occur on an infinitely fast and an infinitely slow time scale, respectively. These limit regimes provide simple, insensitive performance bounds that only depend on appropriately defined load factors, and thus render the detailed statistical characteristics of the system largely irrelevant. For a broad class of Markov-type fading processes, we further proved that the performance varies monotonically with the time scale of the rate variations.

At a qualitative level, the finding that mobility-related rate variations improve performance resembles the generic principle underlying opportunistic scheduling. In the present context, however, the performance improvement does not rely on opportunistic scheduling, but also occurs for example in case of channel-oblivious round-robin scheduling. Instead, informally stated, it arises from the fact that flow-level performance measures behave as convex func-

tions of the rate processes.

From a practical perspective, the quasi-stationary regime provides an accurate conservative estimate of the flow-level performance when the traffic load generated by users with large rate variations is limited. This allows the derivation of simple dimensioning rules *as if* users were static, along the lines [8]. In cases where users with large rate variations account for a significant part of the traffic load, the performance becomes sensitive to the precise traffic and fading statistics. It may then be necessary to take mobility and shadowing effects into account.

Finally, note that the positive impact of mobility relies here on the assumption of perfect rate predictions. This is reflected in the model by the fact that the marginal distributions of the feasible rates do not depend on the time scale of the rate variations. It would be interesting to study the extent to which the estimation and prediction problems due to high-frequency fading offset the performance improvements established in the present paper.

A Proof of Theorem 1

We prove the result for Cox flow size distributions which are known to form a dense subset of the set of all distributions with non-negative support. Specifically, we assume that class-k flows have i.i.d. exponential sizes of mean $1/\mu_k$, and generate a new class-l flow with probability p_{kl} when completed. By creating additional classes and dividing each random flow size into a random number of exponential phases, the model is then sufficiently general to cover any flow size distribution. The total traffic load is given by:

$$\rho := \lambda(I - P)^{-1}(\mu C)^{-1},$$

where $\lambda \equiv (\lambda_k)$ is a row vector, I is the identity matrix, $P \equiv (p_{kl})$, and $\mu \equiv (\mu_k)$ and $C \equiv (C_k)$ are diagonal matrices. Similarly, we assume that the rate process $R_{ki}(t)$ is a function of a finite-state Markov process $\sigma_{ki}(t)$. By increasing the number of states, such a Markov process can approximate any stationary and ergodic process.

The stochastic process $\{N(t), \sigma(t)\}$, where $N(t)$ and $\sigma(t)$ denote the row vectors $(N_k(t))$ and $(\sigma_{ki}(t))$, $k = 1, \ldots, K$, $i = 1, \ldots, N_k(t)$, respectively, is an irreducible Markov process. Define the workload at time t as:

$$W(t) := |N(t)(I - P)^{-1}(\mu C)^{-1}|.$$

For any sequence of initial states $\{N^{(j)}(0), \sigma^{(j)}(0)\}_j$ with

$$\lim_{j\to\infty} \frac{W^{(j)}(0)}{j} = 1, \tag{16}$$

we will prove that the sequence of workload processes $\{W^{(j)}(t)\}_j$ satisfies for any $t < t_0 \equiv 1/(G^* - \rho)$:

$$\lim_{j\to\infty} \frac{\mathrm{E}\left[W^{(j)}(jt)\right]}{j} = 1 - (G^* - \rho)t.$$

As the workload defines a Lyapunov function for the Markov process $\{N(t), \sigma(t)\}$, the proof then follows from Foster's stability criterion [21].

Denote by $A(t)$, $B(t)$ and $D(t)$, respectively, the row vectors of the number of exogenous arrivals, endogenous arrivals and departures of class-k flows up to time t, $k = 1, \ldots, K$. We have:

$$N(t) = N(0) + A(t) + B(t) - D(t). \tag{17}$$

Let $D'(t)$ be the row vector of the maximum number of departures of class-k flows up to time t, thus assuming that these flows are served at rate $S_k \equiv G^* \sup_t R_k(t)$. Similarly, let $B'(t)$ be the row vector of the maximum number of endogenous arrivals of class-k flows up to time t. We get:

$$W(t) \leq W(0) + |(A(t) + B'(t))(I - P)^{-1}(\mu C)^{-1}|. \tag{18}$$

Denoting by e the row vector $(1, \ldots, 1)$, it follows from the strong law of large numbers that:

$$\frac{A(jt)}{j} \overset{\text{a.s.}}{\to} \lambda t, \ \frac{D'(jt)}{j} \overset{\text{a.s.}}{\to} e\mu S_k t, \ \frac{B'(jt)}{j} \overset{\text{a.s.}}{\to} e\mu P S_k t,$$

when $j \to \infty$. In particular, there exists for any sequence of initial states $\{N^{(j)}(0), \sigma^{(j)}(0)\}_j$ satisfying (16) a subsequence denoted by indices j' such that for any $t < t_0$:

$$\frac{W^{(j')}(j't)}{j'} \overset{\text{a.s.}}{\to} \bar{W}(t).$$

The function $\bar{W}(\cdot)$ is usually referred to as a 'fluid limit' due to the time-space scaling [13]. As $D(t) \leq |D'(t)|$ and $B(t) \leq B'(t)$, there exists a subsequence of j', denoted by indices j'', and continuous functions $\bar{B}$ and $\bar{D}$ such that for any $t < t_0$:

$$\frac{B^{(j'')}(j''t)}{j''} \overset{\text{a.s.}}{\to} \bar{B}(t), \ \frac{D^{(j'')}(j''t)}{j''} \overset{\text{a.s.}}{\to} \bar{D}(t) \text{ when } j'' \to \infty.$$

It then follows from (17) that:

$$\frac{N^{(j'')}(j''t)}{j''} \overset{\text{a.s.}}{\to} \bar{N}(t), \text{ when } j'' \to \infty,$$

where $\bar{N}$ is the continuous function defined by:

$$\bar{N}(t) := \bar{N}(0) + \lambda t + \bar{B}(t) - \bar{D}(t). \tag{19}$$

Now as flow sizes are i.i.d. exponential, the number of departures of class-k flows during the time interval $[u,v]$ satisfies:

$$D_k(v) - D_k(u) \leq \sum_{i=1}^{M_k(u,v)} 1\left\{F_{ki} \leq \frac{G^*}{|m(u,v)|}\int_u^v R_{ki}(t)dt\right\},$$

where $M(u,v) := N(u) + A(v) - A(u) + B(v) - B(u)$ and $m(u,v) := N(u) - D(v) + D(u)$ correspond to the row vectors of the maximum and the minimum number of active class-k flows during $[u,v]$, respectively. Analogously, we define $\bar{M}(u,v) := \bar{N}(u) + \lambda(v-u) + \bar{B}(v) - \bar{B}(u)$ and $\bar{m}(u,v) := \bar{N}(u) - \bar{D}(v) + \bar{D}(u)$. For any u such that $|\bar{N}(u)| > 0$, we have $|m(u,v)| > 0$ for sufficiently small $v > u$. For any $\varepsilon > 0$, there exists l such that for all $j'' \geq l$,

$$\frac{1}{j''}|N^{(j'')}(j''u) - D^{(j'')}(j''v) + D^{(j'')}(j''u)| \geq |m(u,v)|(1-\varepsilon) \quad \text{a.s.},$$

and by the ergodicity of the process $R_k(t)$,

$$\Pr\left[\sup_{j\geq l}\left\{\frac{1}{j(v-u)}\int_{ju}^{jv} R_k(t)dt\right\} > C_k(1+\varepsilon)\right] \leq \varepsilon.$$

Writing

$$D_k(v) - D_k(u) \leq$$
$$\sum_{i=1}^{M_k(u,v)} 1\left\{F_{ki} \leq \frac{G^* C_k(1+\varepsilon)(v-u)}{|m(u,v)|}\right\} + \sum_{i=1}^{M_k(u,v)} 1\left\{\frac{1}{v-u}\int_u^v R_{ki}(t)dt > C_k(1+\varepsilon)\right\},$$

it follows from the strong law of large numbers that:

$$\bar{D}_k(v) - \bar{D}_k(u) \leq \bar{M}_k(u,v)\left(\Pr\left[F_k \leq \frac{G^* C_k(1+\varepsilon)(v-u)}{|\bar{m}(u,v)|(1-\varepsilon)}\right] + \varepsilon\right).$$

Since this inequality holds for any $\varepsilon > 0$, we deduce:

$$\bar{D}_k(v) - \bar{D}_k(u) \leq \bar{M}_k(u,v)\Pr\left[F_k \leq \frac{G^* C_k(v-u)}{|\bar{m}(u,v)|}\right].$$

Similarly, using the fact that

$$D_k(v) - D_k(u) \geq \sum_{i=1}^{N_k(u)} 1\left\{F_{ki} \leq \frac{G(|m(u,v)|)}{|M(u,v)|}\int_u^v R_{ki}(t)dt\right\},$$

we obtain:

$$\bar{D}_k(v) - \bar{D}_k(u) \geq \bar{N}_k(u)\Pr\left[F_k \leq \frac{G^* C_k(v-u)}{|\bar{M}(u,v)|}\right].$$

Using the latter inequalities and the fact that $\bar{M}(u,v)$ and $\bar{m}(u,v)$ tend to $\bar{N}(u)$ when v tends to u, we deduce:

$$\frac{d\bar{D}}{dt}(u) = G^*\frac{\bar{N}(u)}{|\bar{N}(u)|}\mu C.$$

Analogously, one may prove that:

$$\frac{d\bar{B}}{dt}(u) = G^* \frac{\bar{N}(u)}{|\bar{N}(u)|}\mu CP.$$

Now it follows from (19) that for any $t < t_0$ such that $|\bar{N}(t)| > 0$:

$$\frac{d\bar{N}}{dt}(t) = \lambda - G^* \frac{\bar{N}(t)}{|\bar{N}(t)|}\mu C(I-P).$$

Using the fact that

$$\bar{W}(t) = |\bar{N}(t)(I-P)^{-1}(\mu C)^{-1}|,$$

we obtain for any $t < t_0$ such that $\bar{W}(t) > 0$:

$$\frac{d\bar{W}}{dt}(t) = \rho - G^*.$$

We know that $\bar{W}(0) = 1$ in view of (16), so that $\bar{W}(t) = 1 - (G^* - \rho)t$ for any $t < t_0$. In particular, the function $\bar{W}(\cdot)$ is uniquely defined for any $t < t_0$, and

$$\frac{W^{(j)}(jt)}{j} \overset{\text{a.s.}}{\to} \bar{W}(t) \quad \text{when } j \to \infty.$$

Finally, the sequence of r.v.'s $\{\frac{\mathrm{E}[W^{(j)}(jt)]}{j}\}_j$ is uniformly integrable in view of (18), so that:

$$\lim_{j\to\infty} \frac{\mathrm{E}\left[W^{(j)}(jt)\right]}{j} = \bar{W}(t).$$

B Proof of Theorem 2

We first prove (4) and (5) for the following slotted system. The interval $(0,t)$ is divided into L slots such that the feasible rate of each flow is constant during each slot and equal to the feasible rate at the beginning of the slot. We also assume that when a flow is present at the beginning of a slot, it remains in the system during the entire slot. The proof of (4) and (5) for a non-slotted system then follows from the fact that for $L = 2^p$, $p \geq 1$, the workload and the number of class-k flows in a slotted system where the feasible rate of a flow during a slot is fixed at its maximum in the slot (respectively, its minimum) converge monotonically to $W(t)$ and $N_k(t)$, respectively, when p tends to ∞.
The proof is based on the notion of supermodular functions (see for instance [22]) and on Lorentz' inequality [26]:

Definition 2 *(Supermodular functions)* $f : \mathbb{R}^n \to \mathbb{R}$ *is supermodular if and only if for all* $x, y \in \mathbb{R}^n$, $f(x \vee y) + f(x \wedge y) \geq f(x) + f(y)$, *where* $(x \vee y)_i = x_i \vee y_i$ *and* $(x \wedge y)_i = x_i \wedge y_i$, *for all* $i = 1, \ldots, n$.

Lemma 1 *(Lorentz' inequality) Let $Z_1, \ldots, Z_n$ be identically distributed r.v.'s. For all supermodular functions $f(\cdot)$,* $\mathrm{E}\left[f(Z_1, \ldots, Z_n)\right] \leq \mathrm{E}\left[f(Z_1, \ldots, Z_1)\right]$.

Consider an arbitrary flow, say flow 1, that arrived at the beginning of slot $l \in \{1, \ldots, L\}$. Assume without loss of generality that flow 1 is of class 1. We fix the arrival process up to slot L, the rate processes of all flows except flow 1, and the sizes of all flows except flow 1. We denote by x_j the feasible rate of flow 1 during slot j, by F_1 the size of flow 1, and by E_1 the expectation with respect to the random variable F_1. Let W_L and $N_{k,L}$ be the workload and the number of class-k flows at the end of slot L, respectively. The key result is:

Lemma 2 *For all $k = 1, \ldots, K$, $\mathrm{E}_1\left[f(W_L)\right]$ and $\mathrm{E}_1\left[g(N_{k,L})\right]$ are supermodular and convex functions of $(x_l, \ldots, x_L)$, for all increasing and convex functions $f(\cdot)$ and all increasing functions $g(\cdot)$.*

Proof Let n_j be the number of flows present during slot j, assuming that flow 1 is present during this slot. Note that these numbers do not depend on $(x_l, \ldots, x_L)$ nor on F_1. Let W_L^1 be the workload due to flow 1 at the end of slot L, i.e., the remaining size of flow 1 divided by C_1. As the transmission rate of flow 1 in slot j is $x_j G(n_j)/n_j$, we have:

$$W_L^1 = \frac{1}{C_1} \max\left(0, F_1 - \sum_{j=l}^{L} x_j \frac{G(n_j)}{n_j}\right),$$

which, composed with an increasing and convex function, is known to be supermodular and convex in $(x_l, \ldots, x_L)$ [5]. Now let w_j and $n_{k,j}$ be the workload and the number of active class-k flows at the end of slot L, respectively, assuming flow 1 leaves the system at the end of slot j. Note that these quantities do not depend on $(x_l, \ldots, x_L)$ nor on F_1. We have:

- If $F_1 \leq x_l G(n_l)/n_l$, then $W_L = w_l$ and $N_{k,L} = n_{k,l}$;
- For $l' = l+1, \ldots, L-1$, if $\sum_{j=l}^{l'-1} x_j G(n_j)/n_j < F_1 \leq \sum_{j=l}^{l'} x_j G(n_j)/n_j$, then $W_L = w_{l'}$ and $N_{k,L} = n_{k,l'}$;
- If $F_1 > \sum_{j=l}^{L-1} x_j G(n_j)/n_j$, then $W_L = w_L + W_L^1$ and $N_{k,L} = n_{k,L}$.

Averaging with respect to the size of flow 1, we obtain for all increasing and convex functions $f(\cdot)$:

$$\begin{aligned}
\mathrm{E}_1\left[f(W_L)\right] &= (f(w_l) - f(w_{l+1}))P_1\left(x_l \frac{G(n_l)}{n_l}\right) + \ldots + (f(w_{L-1}) - f(w_L))P_1\left(\sum_{j=l}^{L-1} x_j \frac{G(n_j)}{n_j}\right) \\
&+ f(w_L)P_1\left(\sum_{j=l}^{L-1} x_j \frac{G(n_j)}{n_j}\right) + \mathrm{E}_1\left[f(w_L + W_L^1) 1_{\left\{F_1 > \sum_{j=l}^{L-1} x_j \frac{G(n_j)}{n_j}\right\}}\right],
\end{aligned}$$

where P_1 denotes the c.d.f. of F_1. Note that the sum of the last two terms in the latter expression is simply equal to $\mathrm{E}_1\left[f(w_L + W_L^1)\right]$, which is a supermodular and convex function of $(x_l, \ldots, x_L)$. In addition, it follows from Assumption 1 that for all $m = l, \ldots, L-1$, the function

$$(x_l, \ldots, x_L) \mapsto -P_1\left(\sum_{j=l}^{m} x_j \frac{G(n_j)}{n_j}\right),$$

as the composition of an affine function and a convex function, is supermodular and convex [5]. As $G(n)/n$ decreases in n, we have $w_l \leq \ldots \leq w_L$, so that $\mathrm{E}_1[f(W_L)]$, as the sum of supermodular and convex functions, is supermodular and convex.
Similarly, we have for all increasing functions $g(\cdot)$:

$$\begin{aligned}\mathrm{E}_1[g(N_{k,L})] &= (g(n_{k,l}) - g(n_{k,l+1}))P_1\left(x_l \frac{G(n_l)}{n_l}\right) + \ldots \\ &+ (g(n_{k,L-1}) - g(n_{k,L}))P_1\left(\sum_{j=1}^{L-1} x_j \frac{G(n_j)}{n_j}\right) + g(n_{k,L}).\end{aligned}$$

As $G(n)/n$ decreases in n, we have $n_{k,l} \leq \ldots \leq n_{k,L}$. Thus $\mathrm{E}_1[g(N_{k,L})]$, as the sum of supermodular and convex functions, is supermodular and convex. □

Now, for any function $f(\cdot)$, we have:

$$\mathrm{E}\left[f(W_L^{\mathrm{fl}1})\right] = \mathrm{E}\left[\mathrm{E}_1[f(W_L)]\left(\mathrm{E}[x_l], \ldots, \mathrm{E}[x_l]\right)\right],$$

$$\mathrm{E}[f(W_L)] = \mathrm{E}\left[\mathrm{E}_1[f(W_L)](x_l, \ldots, x_L)\right],$$

$$\mathrm{E}\left[f(W_L^{\mathrm{qs}1})\right] = \mathrm{E}\left[\mathrm{E}_1[f(W_L)](x_l, \ldots, x_l)\right],$$

where $^{\mathrm{fl}1}$ (respectively, $^{\mathrm{qs}1}$) denotes the fluid (respectively, quasi-stationary) regime with respect to flow 1. Similar relations hold for the number of class-k flows. Using the independence of the rate processes, we deduce from Lemma 2 and the fact that $(\mathrm{E}[x_l], \ldots, \mathrm{E}[x_l]) \leq_{icx} (x_l, \ldots, x_L)$ [4] that for all increasing and convex functions $f(\cdot)$, all increasing functions $g(\cdot)$, and all $k = 1, \ldots, K$:

$$\mathrm{E}\left[f(W_L^{\mathrm{fl}1})\right] \leq \mathrm{E}[f(W_L)],\ \mathrm{E}\left[g(N_{k,L}^{\mathrm{fl}1})\right] \leq \mathrm{E}[g(N_{k,L})],$$

Similarly, using the independence of the rate processes, it follows from Lemma 2 and Lorentz' inequality that:

$$\mathrm{E}[f(W_L)] \leq \mathrm{E}\left[f(W_L^{\mathrm{qs}1})\right],\ \mathrm{E}[g(N_{k,L})] \leq \mathrm{E}\left[g(N_{k,L}^{\mathrm{qs}1})\right].$$

We obtain (4) and (5) by applying successively the same reasoning to an arbitrary set of flows satisfying Assumption 1.

We now prove inequality (6). Let T_L be the time spent by an arbitrary flow in the slotted system up to slot L. We prove exactly as in Lemma 2 that for all increasing functions $g(\cdot)$, $\mathrm{E}_1[g(T_L)]$ is a supermodular and convex function of $(x_1, \ldots, x_L)$. We deduce as above that:

$$\mathrm{E}\left[g(T_L^{\mathrm{fl}})\right] \leq \mathrm{E}\left[g(T_L)\right] \leq \mathrm{E}\left[g(T_L^{\mathrm{qs}})\right],$$

and by letting L tend to ∞:

$$\mathrm{E}\left[g(T(t)^{\mathrm{fl}})\right] \leq \mathrm{E}\left[g(T(t))\right] \leq \mathrm{E}\left[g(T(t)^{\mathrm{qs}})\right],$$

where $T(t)$ denotes the time spent by an arbitrary flow in the non-slotted system up to time t. We obtain (6) by letting t tend to ∞.

C Proof of Theorem 3

The proof is based on the following recent result by Hu & Pan [16] (a similar result had been used by Chang, Chao & Pinedo [11]).

Lemma 3 *Let $\{Z(t), t \in \mathbb{R}\}$ be a process satisfying Assumption 2. For all integers n, all $(r_1, \ldots, r_n)$ and $(s_1, \ldots, s_n)$ such that $r_1 \leq s_1$ and $r_i - r_{i-1} \leq s_i - s_{i-1}$ for $i = 2, \ldots, n$, we have:*

$$\mathrm{E}\left[f(Z(r_1), \ldots, Z(r_n))\right] \geq \mathrm{E}\left[f(Z(s_1), \ldots, Z(s_n))\right]$$

for all supermodular functions $f : \mathbb{R}^n \to \mathbb{R}$ such that the previous expectations exist.

We only give the proof of (10) and (11) (the proof of (12) is then similar to that of (6)). Again, it is sufficient to prove these inequalities for a slotted system. Consider some arbitrary flow 1 satisfying Assumptions 1 and 2. Denote by $x_j = R_1(t \times j/L)$ (respectively, $x_j^{(s)} = R_1(st \times j/L)$) the rate of flow 1 during slot j in the actual system (respectively, when its rate process is accelerated by a factor $s > 1$). It follows from the independence of the rate processes and Lemmas 2 and 3, that for all increasing and convex functions $f(\cdot)$, all increasing functions $g(\cdot)$, and all $k = 1, \ldots, K$:

$$\mathrm{E}\left[f(W_L^{(s)1})\right] \leq \mathrm{E}\left[f(W_L)\right], \ \mathrm{E}\left[g(N_{k,L}^{(s)1})\right] \leq \mathrm{E}\left[g(N_{k,L})\right],$$

where ${}^{(s)1}$ refers to the system where the rate process of flow 1 is accelerated by a factor s. We obtain (10) and (11) by applying successively the same reasoning to an arbitrary set of flows satisfying Assumptions 1 and 2.

References

[1] Agrawal, R., Bedekar, A., La, R.J., Subramanian, V. (2001). Class and channel condition based weighted proportional fair scheduler. In: *Teletraffic Engineering in the Internet Era, Proc. ITC-17*, Salvador da Bahia, eds. J.M. de Souza, N.L.S. da Fonseca, E.A. de Souza e Silva (North-Holland, Amsterdam), 553–565.

[2] Agrawal, R., Subramanian, V. (2002). Optimality of certain channel-aware scheduling policies. In: *Proc. 40th Annual Allerton Conf. Commun. Control., Comp.*, 1532–1541.

[3] Andrews, D.M., Kumaran, K., Ramanan, K., Stolyar, A.L., Vijayakumar, R., Whiting, P.A. (2004). Scheduling in a queueing system with asynchronously varying service rates. *Prob. Eng. Inf. Sc.*, to appear.

[4] Baccelli, F., Brémaud, P. (2003). *Elements of Queueing Theory*, Springer Verlag.

[5] Bäuerle, N. (1997). Inequalities for stochastic models via supermodular orderings. *Commun. Stat. – Stoc. Models* **13**, 181–201.

[6] Bäuerle, N., Rolski, T. (1998). A monotonicity result for the workload in Markovian modulated queues. *J. Appl. Prob.* **35**, 741–747.

[7] Bender, P., Black, P., Grob, M., Padovani, R., Sindhushayana, N., Viterbi, A. (2000). CDMA/HDR: a bandwidth-efficient high-speed wireless data service for nomadic users. *IEEE Commun. Mag.* **38** (7), 70–77.

[8] Bonald, T., Proutière, A. (2003). Wireless downlink channels: user performance and cell dimensioning. In: *Proc. ACM Mobicom 2003*.

[9] Borst, S.C. (2003). User-level performance of channel-aware scheduling algorithms in wireless data networks. In: *Proc. Infocom 2003*.

[10] Borst, S.C., Whiting, P.A. (2003). Dynamic channel-sensitive scheduling algorithms for wireless data throughput optimization. *IEEE Trans. Veh. Techn.* **52**, 569–586.

[11] Chang, C.S., Chao, X., Pinedo, M. (1991). Monotonicity results for queues with doubly stochastic Poisson arrivals: Ross' conjecture. *Adv. Appl. Prob.* **23**, 210–228.

[12] Chaponniere, E.F., Black, P.J., Holtzman, J.M., Tse, D.N.C. (2002). Transmitter directed code division multiple access system using path diversity to equitably maximize throughput. US Patent 6,449,490.

[13] Dai, J.G. (1995). On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid limit models. *Ann. Appl. Prob.* **5**, 49–77.

[14] Grossglauser, M., Tse, D.N.C. (2002). Mobility increases the capacity of ad hoc wireless networks. *IEEE/ACM Trans. Netw.* **10**, 477–486.

[15] Holtzman, J.M. (2000). CDMA forward link waterfilling power control. In: *Proc. IEEE VTC 2000 Spring Conf.*, 1663–1667.

[16] Hu, T., Pan, X. (2000). Comparisons of dependence for stationary Markov processes. *Prob. Eng. Inf. Sc.* **14**, 299–315.

[17] Jalali, A., Padovani, R., Pankaj, R. (2000). Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system. In: *Proc. IEEE VTC 2000 Spring Conf.*, 1854–1858.

[18] Kelly, F.P. (1979). *Reversibility and Stochastic Networks*, Wiley.

[19] Kushner, H.J., Whiting, P.A. (2002). Asymptotic properties of Proportional Fair sharing algorithms. In: *Proc. 40th Annual Allerton Conf. Commun. Control., Comp.*, 1051–1059.

[20] Liu, X., Chong, E.K.P., Shroff, N.B. (2001). A framework for opportunistic scheduling in wireless networks. *Comp. Netw.* **41**, 451–474.

[21] Meyn, S.P., Tweedie, R.L. (1993). *Markov Chains and Stochastic Stability*, Springer Verlag.

[22] Müller, A., Stoyan, D. (2002). *Comparison Methods for Stochastic Models and Risks*, Wiley.

[23] Proutière, A. (2003). Queues in random environment. Work in progress.

[24] Shakkottai, S., Stolyar, A.L. (2001). Scheduling algorithms for a mixture of real-time and non-real time data in HDR. In: *Teletraffic Engineering in the Internet Era, Proc. ITC-17*, Salvador da Bahia, eds. J.M. de Souza, N.L.S. da Fonseca, E.A. de Souza e Silva (North-Holland, Amsterdam), 793–804.

[25] Stolyar, A.L. (2004). On the asymptotic optimality of the gradient scheduling algorithm for multi-user throughput allocation. *Oper. Res.*, to appear.

[26] Tchen, A.H. (1980). Inequalities for distributions with given marginals. *Ann. Appl. Prob.* **8**, 812–827.

[27] Tsibonis, V., Georgiadis, L., Tassiulas, L. (2003). Exploiting wireless channel state information for throughput maximization. In: *Proc. Infocom 2003.*