

# Flow-level performance and capacity of wireless networks with user mobility

Thomas Bonald, Sem Borst, Nidhi Hegde, Matthieu Jonckheere, Alexandre Proutière

▶ **To cite this version:**

Thomas Bonald, Sem Borst, Nidhi Hegde, Matthieu Jonckheere, Alexandre Proutière. Flow-level performance and capacity of wireless networks with user mobility. Queueing Systems, 2009, 10.1007/s11134-009-9144-7 . hal-01244795

**HAL Id: hal-01244795**
**https://hal.science/hal-01244795**

Submitted on 16 Dec 2015

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Flow-Level Performance and Capacity of Wireless Networks with User Mobility

Thomas Bonald*, Sem Borst†,⋆, Nidhi Hegde*, Matthieu Jonckheere⋆, Alexandre Proutiere‡

*Orange Labs, Paris, France

†Bell Labs, Alcatel-Lucent, P.O. Box 636, Murray Hill, NJ 07974, USA

⋆Department of Mathematics & Computer Science, Eindhoven University of Technology
P.O. Box 513, 5600 MB Eindhoven, The Netherlands

‡Microsoft Research, Cambridge, England

September 6, 2009

**Abstract**

The performance evaluation of wireless networks is severely complicated by the specific features of radio communication, such as highly variable channel conditions, interference issues, and possible hand-offs among base stations. The latter elements have no natural counterparts in wireline scenarios, and create a need for novel performance models that account for the impact of these characteristics on the service rates of users.
Motivated by the above issues, we review several models for characterizing the capacity and evaluating the flow-level performance of wireless networks carrying elastic data transfers. We first examine the flow-level performance and stability of a wide family of so-called $\alpha$-fair channel-aware scheduling strategies. We establish that these disciplines provide maximum stability, and describe how the special case of the Proportional Fair policy gives rise to a Processor-Sharing model with a state-dependent service rate. Next we turn attention to a network of several base stations with inter-cell interference. We derive both necessary and sufficient stability conditions and construct lower and upper bounds for the flow-level performance measures. Lastly we investigate the impact of user mobility that occurs on a slow time scale and causes possible hand-offs of active sessions. We show that the mobility tends to increase the capacity region, both in the case of globally optimal scheduling and local $\alpha$-fair scheduling. It is additionally demonstrated that the capacity and user throughput improve with lower values of the fairness index $\alpha$.

## 1 Introduction

The performance evaluation of wireless networks centers on similar metrics as in wireline environments, such as user-perceived throughputs, delay characteristics, and loss rates. However, the evaluation of these performance metrics is severely complicated by the specific features of wireless communications, e.g., uncertain and highly variable channel conditions, interference issues, and possible hand-offs among base stations (BS's) associated with long-range user mobility. The latter features have no natural counterparts in wireline scenarios, and create a strong need to develop and analyze novel models for the performance evaluation of wireless networks. In fact, even the characterization of the network capacity, which is straightforward in

wireline systems with fixed link rates, becomes non-trivial in the presence of channel variations, and even mostly intractable in the case of mutual interference.

In order to develop adequate performance models for wireless networks, it is crucial to identify the primary sources of channel variations. First of all, the channel quality may differ 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. Multi-path fading arises on the level of a wavelength, and occurs on a fast time scale that depends on the carrier frequency and user velocity. Path loss and shadow fading manifest themselves on a more macroscopic level as a result of distance-related attenuation and scattering due to obstacles and terrain conditions, and tend to vary over a longer time scale. Variations in the path loss due to long-range user mobility force hand-offs of active sessions, and cause a dynamic interaction among neighboring BS's. As a further potential source of complex interaction, transmissions tend to be significantly impacted by activity of surrounding BS's because of interference issues.

The above-described channel variations have a critical impact on the instantaneous transmission rates and long-term throughputs. Specifically, the fast channel fluctuations due to multi-path fading, in combination with the relative delay tolerance of elastic data transfers, open up the possibility of scheduling transmissions to the various users when their channel conditions are relatively favorable [43]. This paradigm has triggered a huge interest in so-called channel-aware scheduling strategies as a means to achieve throughput gains for elastic data users [3, 6, 25, 34, 45, 48, 64]. The most prominent exampe of a channel-aware scheduling strategy is the Proportional Fair (PF) policy, which has been widely adopted in commercial systems [5,27,37,64]. The PF policy in fact belongs to a broader class of so-called utility-based schedulers, which may be implemented via simple gradient-based algorithms [1,4,41,44,52,57]. The achievable throughput gains from channel-aware scheduling vary with the channel statistics of the various users as well as the degree of multi-user diversity. As a result, the service rates of the various users depend on the entire user population in a rather intricate fashion. The latter dependence considerably complicates the evaluation of the relevant performance metrics, and renders even the derivation of stability conditions difficult. The slower channel variations due to long-range user mobility offer less scope for channel-aware scheduling, but also pose a major challenge when it comes to evaluating user-perceived throughputs.

Motivated by the above issues, we review in the present paper several models for determining the capacity and assessing the flow-level performance of wireless networks carrying elastic data transfers. In particular, we survey various results originally reported in [7, 8, 17, 18, 21–23, 39]. In the first part of the paper, we examine the flow-level performance and stability of $\alpha$-fair channel-aware scheduling strategies. We describe how under certain assumptions the flow-level performance of the PF scheduling policy may be evaluated by means of a multi-class Processor-Sharing (PS) model where the total service rate varies with the total number of users. The state-dependent service rate accounts for the fact that the throughput gains achieved by channel-aware scheduling increase with the degree of multi-user diversity. The PS model provides explicit formulas for the distribution of the number of active users, mean transfer delays, and blocking probabilities. In particular, the performance is *insensitive*, in the sense that these measures only depend on the statistical characteristics of the system through a readily computed 'load' factor. The notion of 'cell capacity', critical for dimensioning purposes,

can then be defined independently of the detailed properties of the system [13]. Similar PS-type models have been proposed for various kinds of wireless systems [47, 55]. An early paper developing a PS model for a multi-access system is [63].

For general $\alpha$-fair channel-aware schedulers, the evaluation of the flow-level performance involves a multi-dimensional queueing system that does not seem to be tractable. However, we will show that the stability conditions can still be explicitly characterized, and that the family of $\alpha$-fair schedulers in fact provide maximum stability.

In the second part of the paper, we consider a network of several BS's with inter-cell interference, where BS's remain on as long as there are any active users in the corresponding cell, and turn off otherwise. The resulting dynamic interaction among interfering BS's is quite complex and renders an exact analysis elusive in general. In the single-class case, the model reduces to a so-called coupled-processors model, which even for two queues is barely tractable [28, 29, 33], reflecting the complexity of the model in general. Therefore, we focus on the derivation of bounds and approximations. In particular, we derive both necessary and sufficient conditions for stability. We also construct lower and upper bounds for flow-level performance measures, by assuming minimum and maximum interference for either the cell under consideration itself or for all its neighbors.

In the third and final part of the paper, we investigate the impact of user mobility that occurs on a slow time scale and manifests itself in the form of rate variations at flow level. We first consider a scenario where the mobility remains confined to a single cell. Due to these slower rate variations, the above-mentioned insensitivity of the PF strategy is lost, and the performance depends in some complicated fashion on the detailed rate statistics and traffic characteristics. In order to obtain tractable performance estimates, we introduce two limit regimes, termed *fluid* and *quasi-stationary* regime, and use stochastic comparison techniques to show that these yield optimistic and conservative performance estimates, respectively. The latter estimates are particularly useful, as 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.

Next we turn attention to a network of several BS's where the user mobility extends across cells and forces hand-offs of active sessions. We demonstrate that the mobility tends to increase the capacity region, not only in case of globally optimal scheduling, but also when each of the BS's adopts a local $\alpha$-fair discipline. At a qualitative level, the finding that mobility-induced rate variations improve the performance, ties in with the generic rationale for channel-aware scheduling described earlier. It further resonates with the observation in [35] that mobility increases the capacity of ad hoc wireless networks. In the present context, however, the performance improvement does not rely on channel-aware scheduling, but also occurs for example in the 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. In addition, we establish that the capacity and performance improve with lower values of the fairness index $\alpha$. Interestingly enough, in contrast to the situation without user mobility, the overall improvement in capacity and performance is not necessarily at the expense of users in unfavorable conditions.

The remainder of the paper is organized as follows. In Section 2 we present a detailed model description and then proceed to describe how the flow-level performance of the PF scheduling policy may be evaluated by means of a PS model. Next we derive necessary and sufficient conditions for the existence of a scheduling strategy that achieves stability. As a by-product,

we establish that the family of $\alpha$-fair schedulers provide maximum stability. In Section 3 we consider a network of several BS's with inter-cell interference. We examine the impact of mutual interference, establish necessary and sufficient conditions for stability, and construct lower and upper bounds for flow-level performance measures. In Section 4 we investigate the impact of rate variations associated with user mobility on a slower time scale. We first focus on a scenario where the mobility remains confined to a single cell, and prove that two limit regimes yield explicit, insensitive performance bounds. Last, we turn attention to a network of several BS's where the user mobility extends across cells and causes hand-offs of active sessions. We demonstrate that the mobility tends to increase the capacity region, both in case of globally optimal scheduling and in case of a local $\alpha$-fair discipline. It is further shown that the capacity and performance improve with lower values of the fairness index $\alpha$. We make some concluding remarks in Section 5.

# 2 Flow-level performance of channel-aware scheduling strategies

We consider a wireless system carrying elastic traffic from $K$ classes. Each class represents a category of statistically identical users in terms of flow sizes and rate characteristics. Class-$k$ users arrive as a Poisson process of rate $\lambda_k$ (per time unit), and have generally distributed flow sizes $F_k$ (in bits) with mean $\xi_k$. Denote by $\sigma_k := \lambda_k \xi_k$ the offered traffic of class $k$ (in bits per time unit), and define $\sigma := (\sigma_1, \ldots, \sigma_K)$.

As mentioned earlier, channel-aware scheduling causes the set of feasible service rate vectors to depend on the active user population in a fairly intricate fashion. In order to capture that dependence, we define $R(n) \subseteq \mathbb{R}_+^K$ as the set of all feasible service rate vectors for a given user population $n \in \mathbb{N}^K$. The set $R(n)$ has a rather complicated structure in general, but can be characterized through a linear programming formulation. For further details, we refer to [19, 25, 26].

When the user population is $n = (n_1, \ldots, n_K) \in \mathbb{N}^K$, each class-$k$ user receives service at rate $\phi_k(n)/n_k$, with $\phi(n) = (\phi_1(n), \ldots, \phi_K(n)) \in R(n)$ representing the service rate vector for the various classes as function of the user population. The function $\phi(\cdot)$ will frequently be referred to as the allocation function.

Denote by $N(t) = (N_1(t), \ldots, N_K(t))$ the user population at time $t$. Let $(N_1, \ldots, N_K)$ be a random vector representing the number of users of the various classes at an arbitrary epoch in statistical equilibrium (assuming it exists). Denote by $N := N_1 + \cdots + N_K$ the total number of users in the system.

## 2.1 Proportional Fair scheduling

In this subsection, we consider a single-cell-scenario and assume that the allocation function is of the form

$$\phi_k(n) = \bar{R}_k G(n_1 + \cdots + n_K) \frac{n_k}{n_1 + \cdots + n_K}. \tag{1}$$

In this expression, the coefficient $\bar{R}_k$ may be interpreted as the time-average feasible transmission rate of a class-$k$ user (if it were allocated the full transmission resources). The function $G(\cdot)$ captures the throughput gains achieved by channel-aware scheduling, and represents the benefit that each user receives compared to a channel-oblivious round-robin discipline. This

function will be increasing, reflecting the fact that the throughput gains increase with the degree of multi-user diversity. Note that the form of (1) assumes that the relative scheduling gains are identical for all classes and only depend on the user population through the total number of active users. Also, define $G^* := \lim_{n\to\infty} G(n)$.

It may be shown that the allocation function (1) arises in the case of a PF scheduling strategy, assuming the relative rate variations (around the time-average values) of the various user classes to be statistically identical. The latter assumption entails that the instantaneous transmission rate of the $i$-th class-$k$ user $R_{i,k}$ is distributed as $\bar{R}_k Y_{ik}$, where the $Y_{ik}$'s are independent and identically distributed copies of some generic random variable $Y$ with unit mean. Now suppose that the system operates in a time-slotted fashion, with rate variations from slot to slot, and that in each time slot we select the user with the highest instantaneous rate relative to its time-average rate, i.e., the user with the maximum value of $Y_{ik}$. Then each user is equally likely to be selected for service, and given that the $i$-th class-$k$ user is selected, its expected transmission rate is $\mathbb{E}\{R_{ik}|Y_{ik} \geq Y_{jl} \text{ for all } j,l\} = \bar{R}_k \mathbb{E}\{Y_{ik}|Y_{ik} \geq Y_{jl} \text{ for all } j,l\} = \bar{R}_k \mathbb{E}\{\max_{l=1,\ldots,K} \max_{j=1,\ldots,n_k} Y_{jl}\}$. Thus, the expected rate of each class-$k$ user is exactly given by $\phi_k(n)/n_k$ in (1), with $G(n) := \mathbb{E}\{\max\{Y_1,\ldots,Y_n\}\}$ and $Y_1,\ldots,Y_n$ independent and identically distributed copies of the random variable $Y$. The assumption that the relative rate variations are statistically identical (and in fact exponentially distributed), is roughly valid when the users for example have Rayleigh fading channels and the feasible rates are approximately linear in the SNR (signal-to-noise ratio). The latter approximation is reasonably accurate when the SNR is low, and then yields $G(M) = \sum_{m=1}^{M} 1/m$.

In order to see that the allocation vector $\phi(n)$ in (1) is proportional fair, observe that the set $R(n)$ of all feasible service rate vectors is still complicated, even under the above symmetry assumption. However, each achievable throughput vector $T(n) \in R(n)$ satisfies $\sum_{k=1}^{K} \sum_{i=1}^{n_k} T_{ik}(n)/\bar{R}_k \leq \mathbb{E}\{\max_{k=1,\ldots,K} \max_{i=1,\ldots,n_k} R_{ik}/\bar{R}_k\} = \mathbb{E}\{\max_{k=1,\ldots,K} \max_{i=1,\ldots,n_k} Y_{ik}\} = G(n_1 + \cdots + n_K)$. Hence, each achievable throughput vector satisfies $\sum_{k=1}^{K} n_k \sum_{i=1}^{n_k} T_{ik}(n)/\phi_k(n) \leq n_1 + \cdots + n_K$, which means that the allocation vector $\phi(n)$ is proportional fair. For further details, we refer to [17, 18].

We now proceed to show that in case the allocation function is of the form (1), one can explicitly evaluate the flow-level performance in terms of the number of active users, mean transfer delays, and blocking probabilities. Further to the earlier model description, we include admission control, and assume that at most $M$ users are admitted in the system simultaneously (possibly $M = \infty$). Users which initiate service requests when there are already $M$ transfers in progress are denied access and abandon. For convenience, let $B_k := F_k/\bar{R}_k$ be the normalized service requirement of a class-$k$ user with mean $\beta_k := \xi_k/\bar{R}_k$. Note that the normalized service requirement encapsulates both the transfer amount (in bits) and the mean transmission rate of a user, and is measured in transmission time rather than data volume. Define $\rho_k := \lambda_k \beta_k = \sigma_k/\bar{R}_k$ as the normalized traffic intensity of class $k$, and by $\rho := \sum_{k=1}^{K} \rho_k$ the total normalized traffic intensity. Let $B_k^r$ be a random variable representing the residual lifetime of $B_k$ and $B_k^r(\cdot)$ the corresponding distribution function, i.e., $B_k^r(x) := \mathbb{P}\{B_k^r < x\} := \frac{1}{\beta_k} \int_{y=0}^{x} \mathbb{P}\{B_k > y\} dy$.

Given that there are $n_k$ class-$k$ users in the system, let $B_{k,i}^r$ be the remaining normalized service requirement of the $i$-th class-$k$ user, $i = 1,\ldots,n_k$, $k = 1,\ldots,K$.

Now observe that the form of the allocation function in (1) implies that the normalized remaining service requirement of each user is reduced at rate $G(n)/n$, which means that the normalized remaining service requirements evolve in a similar probabilistic fashion as the remaining ser-

vice requirements in a multi-class Processor-Sharing (PS) system with arrival rates $\lambda_k$, generic service requirements $B_k$, and service rate $G(n)$ when there are $n$ users in total present. The next proposition follows from well-known results for such a system [30, 40].

**Proposition 2.1** *The PF strategy achieves stability for $\rho < G^*$ or $M < \infty$, in which case*

$$\mathbb{P}\{N_k = n_k, B^r_{k,i} \leq t_{k,i}; i = 1, \ldots, n_k, k = 1, \ldots, K\} = H^{-1} \frac{n! \rho^n}{\phi(n)} \prod_{k=1}^{K} \frac{1}{n_k!} \left(\frac{\rho_k}{\rho}\right)^{n_k} \prod_{i=1}^{n_k} B^r_k(t_{k,i}),$$

*with $n := \sum_{k=1}^{K} n_k \leq M$, $\phi(n) := \prod_{i=1}^{n} G(i)$, and normalization constant $H := \sum_{n=0}^{M} \frac{\rho^n}{\phi(n)}$. In particular,*

$$\mathbb{P}\{N = n\} = H^{-1} \frac{\rho^n}{\phi(n)},$$

*and the blocking probability is given by $L = \mathbb{P}\{N = M\}$.*

Using Little's law, we find that the mean transfer delay experienced by a class-$k$ user is given by

$$\mathbb{E}\{S_k\} = \frac{\beta_k}{\rho(1-L)} \mathbb{E}\{N\}.$$

The above formula reflects the celebrated insensitivity property of the PS discipline, which shows that the mean delay of a class-$k$ user only depends on the service requirement distribution of class $k$ through its mean $\beta_k$. In fact, it may be shown that the conditional expected delay of any user with actual service requirement $b$ is given by

$$\mathbb{E}\{S|B = b\} = \frac{b}{\rho(1-L)} \mathbb{E}\{N\}.$$

Thus, the expected transfer delay incurred by a user is proportional to its normalized service requirement, with factor of proportionality $\mathbb{E}\{N\}/(\rho(1-L))$. The latter property embodies a certain fairness principle, which means that users with larger service requirements tend to experience longer delays. Recall that the normalized service requirement encapsulates both the transfer volume and the mean transmission rate of a user, and is expressed in time units rather than data bits.

**Remark 2.1** Proposition 2.1 extends to the case where users generate sessions consisting of multiple transfer requests separated by random 'think times' as in [16]. In that case, the traffic intensity should be calculated so as to include the mean number of transfer requests per session.

**Remark 2.2** We refer to [20] for an extension of the model to an integrated system supporting a mixture of elastic flows and adaptive streaming traffic as considered in [14, 42] in a wireline setting.

## 2.2 Generic stability conditions

We now examine under what conditions an allocation function $\phi(\cdot)$ exists, with $\phi(n) \in R(n)$ for all $n \in \mathbb{N}^K$ for given sets $R(n)$, such that the system is stable. We borrow from the results originally reported in [23], and use similar ideas as developed by Massoulié [50] to incorporate

load balancing as an additional control mechanism. In Section 3 we will investigate under what conditions the system is stable for a given allocation function $\phi(\cdot)$, which turns out to be a harder problem.

Henceforth we assume exponentially distributed flow sizes with unit mean. (The latter assumption does not involve any loss of generality as the sets $R(n)$ can easily be scaled to account for different exponential service rates. While in some cases the stability results are conjectured to hold for general flow size distributions, such an extension entails major technical difficulties in the proofs, and there are also cases where the stability condition is likely to be sensitive to the flow size distribution.) Thus the process $N(t)$ tracking the active user population is a $K$-dimensional birth-death process with birth rates $\lambda_i$ and death rates $\phi_i(N(t))$. In particular, stability of the system corresponds to positive recurrence of the latter process.

We make two natural assumptions concerning the sets $R(n)$ which will play a crucial role in deriving the stability conditions. First of all, each of the sets $R(n)$ is assumed to be convex. Second, the sets $R(n)$ are assumed to be monotone increasing in the user population i.e., if $m \leq n$, then $R(m) \subseteq R(n)$.

The above two assumptions are satisfied in scenarios with globally scheduled medium access control. In these scenarios any convex combination of rate vectors is achievable through time sharing, and additional users may simply be excluded from service without affecting the feasible service rates of the remaining users, ensuring monotonicity.

Scheduled medium access control is commonly used on the downlink of a cellular system, and is by definition 'global' in nature if we restrict attention to a single-cell scenario. However, global scheduling is not always a viable option in multi-cell scenarios where individual BS's tend to make local scheduling decisions and in particular remain on as long as there are any active users to be served. In that case, the stability conditions become far more complicated and delicate, as we will see in Section 3. Also, in the absence of a centralized control entity, medium access is commonly governed by distributed and possibly randomized mechanisms. In those cases, the convexity property may not be satisfied, and the stability conditions entail major complications [15, 49].

Define $R^* \subseteq \mathbb{R}_+^K$ as the closure of $\bigcup_{n \in \mathbb{N}^K} R(n)$, which inherits the convexity of the sets $R(n)$. While the set $R^*$ may have a complicated structure in general, it has a rather simple form in the special case where only a single user is served in each time slot. Denote by $R_k^*$ the maximum possible value of the rate of class-$k$ users (possibly $R_k^* = \infty$). Then $R^* = \text{conv}(\{R_1^* e_1, \ldots, R_K^* e_K\}) = \{x \in \mathbb{R}_+^K : \sum_{k=1}^K \frac{x_k}{R_k^*} \leq 1\}$. If in addition the relative rate variations around the time-average values are statistically identical for all classes, then $R_k^* = \bar{R}_k G^*$ with $G^* := \lim_{n \to \infty} G(n)$ as defined in Section 2.1. In that case, $R^* = \{x \in \mathbb{R}_+^K : \sum_{k=1}^K \frac{x_k}{\bar{R}_k} \leq G^*\}$.

In this subsection, we allow for load balancing as an additional control mechanism, which is modeled through a function $\lambda(n) \in Q$ describing how the arrival rate vector is governed by the user population, with $Q \subseteq \mathbb{R}_+^K$ some given closed convex set. In the absence of load balancing, the set $Q$ is simply a singleton. Such a load balancing strategy is particularly relevant in networks with several BS's where flows along the border between two cells may be assigned to either serving BS (with two 'artificial' classes representing the two options, and a sum constraint on the two arrival rates). Indeed, all the results in the present section apply for networks with several BS's. However, they do rely on the assumption that the allocation vector $\phi(\cdot)$ is a function of the entire user population, which requires some global mechanism

that may be harder to implement in a network with a large number of BS's than in a single isolated cell.
The next proposition states a sufficient as well as a necessary condition for the existence of a combined load balancing strategy $\lambda(\cdot)$ and allocation function $\phi(\cdot)$ that achieve stability.

**Proposition 2.2** *If there exists a pair $(q^*, r^*) \in Q \times R^*$ such that $q^* < r^*$, i.e., $q_i^* < r_i^*$ for all $i = 1, \ldots, K$, then there exist a combined load balancing strategy $\lambda(\cdot)$ and allocation function $\phi(\cdot)$ that achieve stability. If on the other hand $Q \cap R^* = \emptyset$, then stability cannot be achieved.*

**Proof** We start with the proof of the first assertion, which follows along similar lines as in [50] and [23]. Consider the load balancing strategy / allocation function defined by

$$(\lambda^*(n), \phi^*(n)) = \arg \max_{(q,r) \in Q \times R^*} \langle r - q, n \rangle.$$

The above-mentioned properties of the sets $R(n)$ imply that there exists a sequence $\epsilon(n)$ such that the load balancing strategy / allocation function $(\lambda(n), \phi(n)) = (\lambda^*, \phi^*)(n) - \epsilon(n) \in Q \times R(n)$ for all $n \in \mathbb{N}^K$, with $\epsilon(n) \to 0$ as $|n| \to \infty$.
Define the Lyapunov function $F(n) := \max_{(q,r) \in Q \times R^*} \langle r - q, n \rangle$. Denote by $\Delta g(n) := \sum_p q(n,p)(g(p) - g(n))$ the drift of a function $g(\cdot)$ of a Markov process with transition rates $q(\cdot, \cdot)$. Let $\delta > 0$ be fixed. Because of the 1-homogeneity of the function $F(\cdot)$, there exists an $m$ such that $|n| > m$ implies $\Delta F(n) \leq \langle \text{grad} F(n), \Delta n \rangle + \delta$. Noting that $Q \times R^*$ is convex, we obtain

$$\Delta F(n) \leq \langle \arg \max_{(q,r) \in Q \times R^*} \langle r - q, n \rangle, \lambda(n) - \phi(n) \rangle + \delta,$$

and the fact that $\epsilon(n) \to 0$ as $|n| \to \infty$ implies

$$\Delta F(n) \leq -|\lambda^*(n) - \phi^*(n)|^2 + 2\delta \leq -M,$$

for $m$ large enough and $M > 0$.
It remains to be shown that $F(n)$ diverges to infinity when $|n| \to \infty$. For $|n|$ large enough, if there exists a pair $(q^*, r^*) \in Q \times R^*$ such that $q^* < r^*$, then there exists a constant $c$ such that $\max_{(q,r) \in Q \times R^*} \langle \phi - \lambda, n \rangle \geq \langle r^* - q^*, n \rangle \geq c|n|$, which shows that $F(\cdot)$ is divergent. The stability then follows from the Lyapunov-Foster criterion.
The converse statement follows from the simple observation that the long-term mean arrival rate vector and long-term mean service rate vector must be contained in $\text{conv}(Q) = Q$ and $\text{conv}(R^*) = R^*$, respectively. Thus, $Q \cap R^* = \emptyset$ precludes stability.

□

**Remark 2.3** *Note that if there exists no pair $(q^*, r^*) \in Q \times R^*$ such that $q^* < r^*$, then either $Q$ only intersects with the Pareto boundary of $R^*$ or $Q \cap R^* = \emptyset$. Thus the necessary condition established in the above proposition is in fact 'nearly' sufficient for the existence of a load balancing strategy and allocation function that achieve stability.*

In the absence of load balancing, the set $Q$ is simply a singleton, and we obtain the following corollary.

**Corollary 2.1** *If $\lambda \in \text{int}(R^*)$, then there exists an allocation function that achieves stability. If on the other hand, $\lambda \notin R^*$, then stability cannot be achieved.*

The proof of Proposition 2.2 in fact identifies a specific load balancing strategy and allocation function that achieve stability under the sufficient condition (and hence 'nearly always' when the necessary condition is satisfied). The rationale for these is that they maximize the drift of the process $N(t)$ towards the origin at all times.

### 2.3 Maximum stability of $\alpha$-fair schedulers

Similar arguments may be used to study the stability of a broad range of allocation functions $\phi_\alpha(\cdot)$ that correspond to the family of weighted $\alpha$-fair utility-based schedulers [1, 41, 52, 57]. Specifically, define $(\lambda^*(n), \phi^*_\alpha(n)) := \arg\max_{(a,b)\in G(Q)\times G(R^*)}\langle b-a, wn^\alpha\rangle$, and $F(n) := \langle \mathbf{1}, w\frac{n^{\alpha+1}}{1+\alpha}\rangle$, with $\alpha > 0$ and $w \in \mathbb{R}^K_+$ a positive weight vector, $wn^\alpha = (w_1 n_1^\alpha, \ldots, w_K n_K^\alpha)$, $\mathbf{1} = (1, \ldots, 1)$ and $G : \mathbb{R}_+ \to \mathbb{R}$ an increasing and concave function.
First observe that $F(\cdot)$ is $(\alpha+1)$-homogeneous, yielding the upper bound $\Delta F(n) \leq \langle \text{grad}F(n), \Delta n\rangle + o(|n^\alpha|)$ for large $n$. Then note that $\langle \text{grad}F(n), \Delta n\rangle = -\max_{(a,b)\in G(Q)\times G(R^*)}\langle b-a, wn^\alpha\rangle \leq -M$ given the stability condition presented in Proposition 2.2, and the proof arguments may be readily extended.
In particular, in the absence of load balancing, taking $G(x) = \frac{x^{1-\alpha}}{1-\alpha}$, we obtain

$$\phi^*(n) = \arg\max_{b\in G(R^*)}\langle b, wn^\alpha\rangle = \arg\max_{b\in R^*}\langle \frac{b^{1-\alpha}}{1-\alpha}, wn^\alpha\rangle,$$

which corresponds to weighted $\alpha$-fair utility functions $U(x) = \frac{x^{1-\alpha}}{1-\alpha}$ for $\alpha > 0$, with the convention that $U(x) = G(x) = \log(x)$ for $\alpha = 1$. The latter family of utility functions covers the most common fairness notions, such as proportional fairness ($\alpha = 1$), and max-min fairness ($\alpha = \infty$). Thus, we conclude that the family of $\alpha$-fair utility-based schedulers with $\alpha > 0$ achieve stability under the sufficient condition (and therefore 'nearly always' when the necessary condition is satisfied). This result is in the same spirit as in [11, 12, 46], where the rate region is however fixed and does not depend on the user population.
As described earlier, in the special case where only a single user is served in each time slot, we have $R^* = \{x \in \mathbb{R}^K_+ : \sum_{k=1}^K \frac{x_k}{R^*_k} \leq 1\}$, and the sufficient stability condition $\lambda \in \text{int}(R^*)$ reduces to $\sum_{k=1}^K \lambda_k/R^*_k < 1$. Note that this corresponds to the stability condition for a work-conserving single-server system where class-$k$ users can always be served at rate $R^*_k$. This somewhat surprising fact may be explained by the observation that under a weighted $\alpha$-fair strategy every class will either be served at the maximum possible rate ionr not at all whenever any of them is unstable.
It is interesting to observe that the above results contrast with the fact that utility-based scheduling strategies generally fail to provide maximum stability guarantees at *packet* level, see for instance [2, 54]. Various simple queue-length-based strategies on the other hand do achieve stability at packet level whenever possible [3, 60, 61]. In order to reconcile these paradoxical facts, it is worth observing that while such utility-based strategies operate agnostically of the queue lengths at packet level, they *do* respond to congestion that occurs at flow level. Thus, from a stability perspective, the behavior of a utility-based strategy at flow level shows resemblance to that of a queue-length-based strategy at packet level. An important related finding in the context of 'imperfect' scheduling in multi-hop networks is described in [45]. However, a crucial distinction is that at *packet* level channel fluctuations give rise to *random* time-varying service rates for the various users, which are *independent* of the number of packets stored in the buffer. In contrast, the feasible service rates for the various classes at *flow* level are *deterministic* as the channel fluctuations 'average out', but they *vary* with the number of users because the scheduling gains increase with the degree of multi-user diversity as mentioned earlier.

# 3 Networks with inter-cell interference

In the present section we examine the flow-level stability and performance of networks with several cells subject to interference between BS's. The dynamics of such systems are quite complex since the activity state of each BS affects the service rates of users in neighboring cells, which in turn influences the activity state of the corresponding BS's. We model the system as a processor-sharing network where the service rate of each class depends on the number of active users in the other cells. In order to obtain more tractable results, we assume that there is only one class of users per cell.
We first examine under what conditions the system is stable for a given allocation function, which turns out to be a much harder problem than the one considered in the previous section. We mostly borrow from the results originally reported in [24, 38]. Earlier results of this type were obtained by Szpankowski [58, 59]. Similar problems were also recently studied by Hansen *et al.* [36].
As the stability conditions turn out to be complicated and difficult to calculate in general, we derive bounds that can be easily evaluated and do not depend on detailed statistical characteristics of the system. We then give approximations for key performance metrics like the number of active users, transfer delays and user throughputs.

Throughout the section we continue to assume exponentially distributed flow sizes with unit mean, so that stability corresponds to positive recurrence of the Markov process $N(t)$ representing the active user population. In case the process $N(t)$ is not positive-recurrent, a restricted version $(N_k(t))_{k\in\mathcal{L}}$, $\mathcal{L} \subseteq \{1, \ldots, K\}$, may still be 'stable' in a certain sense. Such a restricted version will however not be a Markov process in general, and the notion of positive recurrence may not readily apply. Therefore the process $(N_k(t))_{k\in\mathcal{L}}$ will be called 'stable' if for any $\epsilon > 0$, there exists a finite set $S_\epsilon$ such that

$$\mathbb{P}((N_k(t))_{k\in\mathcal{L}} \notin S_\epsilon) \leq \epsilon \quad \text{for all } t,$$

and otherwise the process is said to be *unstable*.

## 3.1 Partially decreasing service allocations

In wireless networks, the feasible service rates at a given BS typically decrease in a complex way when neighboring BS's are active due to mutual interference. As a result, the service rate of a given class will usually decrease with the number of active users of competing classes in other cells. Motivated by the above observations, we will assume that the allocation function satisfies a natural monotonicity property. Specifically, the allocation function $\phi(\cdot)$ is said to be *partially decreasing* if for all $i$,

$$\phi_i(n) \geq \phi_i(m) \text{ for all } n \leq m \text{ such that } n_i = m_i.$$

Implicitly we assume here that a BS is always on as long as there are any active users in the cell. (The interference between cells in fact provides a potential incentive to turn off BS's, even when there are users to serve, and coordinate the activity patterns of interfering BS's, see for instance [9, 10].) We do not make any further assumptions on the specific form of the allocation function, which tends to be quite intricate and strongly depends on the particular properties of the channel-aware scheduling policy, the fading behavior and the propagation characteristics.

Note that any component-wise decreasing function is partially decreasing, and any function $\phi(\cdot)$ such that $\phi_i(n)$ only depends on $n_i$, is also partially decreasing. Importantly, a multi-class birth-death process with constant birth rates and bounded state-dependent death rates $\phi_i(\cdot)$ is monotone if and only if the function $\phi(\cdot)$ is partially decreasing. Recall that a continuous-time Markov process $N(t)$ is said to be monotone if $\mathbb{E}\{f(N(t))|N(0) = n\}$ is increasing for all $t$ in the initial state $n$ for any bounded increasing function $f(\cdot)$.
The fact that the allocation function $\phi(\cdot)$ is partially decreasing, allows us to establish stability by inductively comparing the process with decoupled versions of it, and determining stability conditions for each of the components.
Define the death rates $\ell^l\phi_i$ by the lower partial limits

$$\ell^l\phi_i(n_1,\ldots,n_l) := \lim_{r\to\infty} \inf_{n_{l+1},\ldots,n_K>r} \phi_i(n_1,\ldots,n_K). \tag{2}$$

The quantity $\ell^l\phi_i(n_1,\ldots,n_l)$ represents the asymptotically worst-case service rate received by class $i$ in a partially saturated system where the numbers of users of classes $l+1,\ldots,K$ tend to infinity. Let $Y^l(t)$ be an $l$-dimensional birth-death process with birth rates $\lambda_i$ and death rates $\ell^l\phi_i$, which may intuitively be interpreted as a partially saturated version of the process $N(t)$, where classes $1,\ldots,l$ are allocated the asymptotically worst-case service rates. Also, define

$$L_i^l(\lambda_1,\ldots,\lambda_l;\phi) := \sum_{n\in\mathbb{N}^l} \ell^l\phi_i(n)\pi^l(n),$$

if $Y^l(t)$ has a unique stationary distribution $\pi^l$, and set $L_i^l(\lambda_1,\ldots,\lambda_l;\phi) := 0$ otherwise. The quantity $L_i^l(\lambda_1,\ldots,\lambda_l;\phi)$ represents the worst-case average service rate received by class $i$ in a partially saturated system where the numbers of users of classes $l+1,\ldots,K$ tend to infinity. For notational convenience, denote $L_i^l(\lambda_1,\ldots,\lambda_l;\phi) := \ell^0\phi_i$ for $l = 0$.

**Proposition 3.1** *Assume the allocation function $\phi(\cdot)$ is bounded and partially decreasing and that there exists an $l$ such that*

$$\lambda_i < L_i^{i-1}(\lambda_1,\ldots,\lambda_{i-1};\phi) \tag{3}$$

*for all $i = 1,\ldots,l$. Then each of the processes $N_1(t),\ldots,N_l(t)$ is stable, regardless of the initial state.*

**Remark 3.1** *Because of the partial monotonicity of $\phi(\cdot)$, the sequence $j \to L_i^j(\lambda_1,\ldots,\lambda_{j-1};\phi)$ is increasing. Hence, if the network is completely symmetric, in the sense that the service rates and the arrival rates of all classes are equal, then the stability region boils down to $\bar{\lambda} \equiv \lambda_i < \ell^0\phi_i \equiv \bar{\phi}$ for all $i$.*

### 3.2 Partially decreasing service allocations with uniform limits

In the previous subsection we made the assumption that the allocation function $\phi(\cdot)$ is partially decreasing. In this subsection, we additionally impose the assumption that the allocation function has uniform limits as the numbers of users of some of the classes tend to infinity. Specifically, an allocation function $\phi(\cdot)$ is said to have *uniform limits at infinity* if for all $i$:

1. There exists a constant $\phi_i^0$ such that $\sup_{n\in\mathbb{N}^K:n_1,\ldots,n_K>r} |\phi_i(n) - \phi_i^0| \to 0$ as $r \to \infty$.

2. For any $k = 1, \ldots, K-1$ and any permutation $\sigma$ on $\{1, \ldots, K\}$, there exists a function $\phi_i^{l,\sigma} : \mathbb{N}^l \to \mathbb{R}$ such that $\sup_{n \in \mathbb{N}^K : n_{\sigma(k+1)}, \ldots, n_{\sigma(K)} > r} |\phi_i(n) - \phi_i^{l,\sigma}(n_{\sigma(1)}, \ldots, n_{\sigma(l)})| \to 0$ as $r \to \infty$.

For example, if the allocation function $\phi(\cdot)$ is of the form $\phi_i(n) = g_i(n_i)h(n)$, where $g_i(\cdot)$ has a limit at infinity, while $h(\cdot)$ is a (component-wise) decreasing function accounting for the mutual interference, then $\phi_i(\cdot)$ has a uniform limit. If $\phi(\cdot)$ has uniform limits, then the partial lower limits $\ell^l\phi$ defined in (2) become true limits in the sense that $\phi(n_1, \ldots, n_K) \to \ell^l\phi(n_1, \ldots, n_l)$ uniformly over $n_1, \ldots, n_l$ as $\min\{n_{l+1}, \ldots, n_K\} \to \infty$.

**Proposition 3.2** *Assume the allocation function $\phi(\cdot)$ is bounded and partially decreasing with uniform limits at infinity. Assume that there exists an index $l$ such that*

$$\begin{aligned} \lambda_i &< L_i^{i-1}(\lambda_1, \ldots, \lambda_{i-1}; \phi) \text{ for all } i \leq l, \\ \lambda_i &> L_i^{l}(\lambda_1, \ldots, \lambda_{l}; \phi) \text{ for all } i > l. \end{aligned}$$

*Then the process $(N_{l+1}(t), \ldots, N_K(t))$ is unstable, regardless of the initial state.*

Applying Proposition 3.1 to all possible permutations of the classes yields sufficient conditions for the global stability of the system, i.e., the positive recurrence of the Markov process $N(t)$. Proposition 3.2 demonstrates that these conditions are also 'nearly' necessary: the system is unstable outside the closure of the set defined by the sufficient conditions (3).

The above propositions show that the stability of a system with $K$ user classes can be expressed in terms of the stationary distributions of a reduced system with $K-1$ classes. These stationary distributions might be sensitive to subtle properties of the allocation function $\phi(\cdot)$, which illuminates the difficulty of characterizing the exact stability region for heterogeneous networks. Consider for example a two-cell network with allocation function $\phi_i(x) = g_i(x_i)h_i(x_j)$. This particular form of allocation function arises in case of two interfering BS's operating according to a channel-aware scheduling discipline. The functions $h_i(\cdot)$ capture the interference between the two BS's, and the functions $g_i(\cdot)$ reflect the scheduling gain, which increases with the number of users due to multiuser diversity. Figure 1 shows the service rate vectors for various numbers of users in each cell, and the frontier of the stability region for two types of allocation functions $\phi(\cdot)$ and $\psi(\cdot)$ given by:

$$\begin{aligned} \phi_i(n) &= \min(3, \log(1+n_i))(1_{n_j>0} + 0.5 \cdot 1_{n_j>0}), \quad j \neq i, \\ \psi_i(n) &= \min(3, \log(1+n_i))\frac{1}{2-(1+j)^{-0.4}}, \quad j \neq i. \end{aligned}$$

In the first scenario, the interference reduces the service rates by a factor 2 as soon as the number of users in the interfering cell is strictly positive. In the second case, the interference is smoother and the impact on the service rates increases with the number of active users in the interfering cell.
The resulting stability regions are not convex and depend in a complicated way on the tradeoff between interference and the gain from channel-aware scheduling. Note that operating the network in the most symmetric fashion (equal arrival rates in each cell) is optimal in the first scenario but not in the second case where the arrival rates maximizing the total traffic that the network can support are non-trivial.

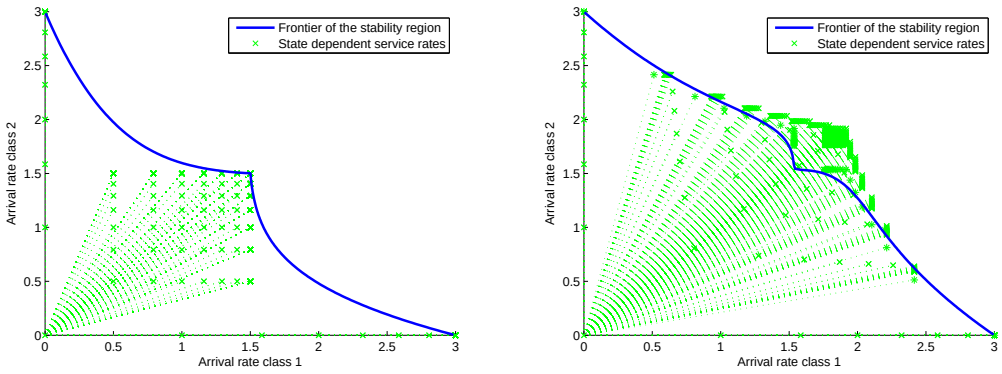



Figure 1: State-dependent service rates and the corresponding stability region.

With an arbitrary number of $K$ user classes, the global stability region involves $K!$ sets of conditions in general. However, depending on the topology and possible symmetries, the number of conditions can possibly be reduced. Consider for example a 3-cell linear network. Suppose that cell 1 is in the middle and interferes with cells 2 and 3, while cells 2 and 3 interfere only with cell 1. To simplify notation, denote by $\phi_i^{\mathcal{A}}$ the service rate received by cell $i$ when cells outside $\mathcal{A} \subseteq \{1,2,3\}$ are supposed to be saturated and cells in $\mathcal{A}$ are supposed to be stationary. The stability conditions boil down to:

$$\lambda_1 < \phi_1^{\emptyset},\ \lambda_2 < \phi_2^{\{1\}},\ \lambda_3 < \phi_3^{\{1\}},$$
$$\text{or } \lambda_2 < \phi_2^{\emptyset},\ \lambda_1 < \phi_1^{\{2\}},\ \lambda_3 < \phi_3^{\{1\}},$$
$$\text{or } \lambda_2 < \phi_1^{\emptyset},\ \lambda_3 < \phi_3^{\emptyset},\ \lambda_1 < \phi_1^{\{2,3\}},$$
$$\text{or } \lambda_3 < \phi_3^{\emptyset},\ \lambda_1 < \phi_1^{\{3\}},\ \lambda_2 < \phi_2^{\{1\}}.$$

The proof of Propositions 3.1 and 3.2 developed in [24] relies both on stochastic comparisons and martingale arguments, and does not involve specific Lyapunov functions. The knowledge of a Lyapunov function provides however valuable insight into the type and speed of convergence of the process to its stationary distribution [51]. Following [39], we now give a construction of a Lyapunov function in the simplest case of two user classes. The construction sheds some light on the difficulty of making such functions explicit, as they should depend on the specific properties of the allocation function $\phi(\cdot)$. The construction of the Lyapunov function relies on solving a Poisson equation. For that purpose, suppose without loss of generality that $\lambda_i + \phi_i(n) < 1$, $i = 1, 2$, for all $n \in \mathbb{N}^2$ and define $P$ as the kernel of a one-dimensional Markov process with transition rates

$$p(n_1, n_1+1) = \lambda_1,\ p(n_1, n_1-1) = \ell^1\phi_1(n_1),\ p(n_1, n_1) = 1 - (\lambda_1 + \ell^1\phi_1(n_1)).$$

**Proposition 3.3** *If*

$$\lambda_1 < \ell^0\phi_1 \text{ and } \lambda_2 < L_2^1(\lambda_1),$$

*then a Lyapunov function for the system is given by*

$$F(n) := \psi^{n_1} + \gamma[n_2 + V(n_1)].$$

*with $\psi > 1$, $\gamma > 0$ some constants to be chosen and $V(\cdot)$ a bounded function defined as the solution of the Poisson equation $(I - P)V = \lambda_2 - \ell^1\phi_2 + \epsilon$, with $\epsilon := L_2^1(\lambda_1) - \lambda_2$.*

### 3.3 Bounds for the stability conditions and performance

In this subsection, we use further stochastic comparisons to derive bounds for the stability conditions and the performance that are far simpler to compute. Recall that $Y^l(t)$ is the $l$-dimensional birth-death process with birth rates $\lambda_i$ and death rates $\ell^l\phi_i$, $i = 1, \ldots, l$. The partial monotonicity of the allocation function implies that

$$(N_1, \ldots, N_l) \leq_{st} (Y_1^l, \ldots, Y_l^l).$$

This stochastic comparison has yielded the stability conditions in Proposition 3.1. However, the stability conditions obtained can be difficult to compute, since they involve the calculation of the stationary distribution $\pi^l$ of the $l$-dimensional process $Y^l(t)$, which does not have a closed form in general unless $l = 1$. Using the fact that $\ell^l\phi_i$ is partially decreasing, $i = 1, \ldots, l$, however, we can derive a looser stochastic comparison, namely

$$(N_1, \ldots, N_l) \leq_{st} (Y_1^l, \ldots, Y_l^l) \leq_{st} (\tilde{Y}_1^{l-1}, \ldots, \tilde{Y}_{l-1}^{l-1}, \tilde{N}_l), \tag{4}$$

where the process $(\tilde{Y}^{l-1}, \tilde{N}_l)$ has death rates $\phi_i^i(n_i)$, $i = 1, \ldots, l-1$, and $\ell^l\phi_l$, with

$$\phi_i^i(n_i) = \lim_{r\to\infty} \inf_{n_j > r, j\neq i} \phi_i(n_1, \ldots, n_K).$$

Note that with these modified death rates, the processes $\tilde{Y}^j = (\tilde{Y}_1^j, \ldots, \tilde{Y}_j^j)$ are themselves Markov processes for all $j$. Moreover, all their components decouple, leading to the following stationary distribution

$$\tilde{\pi}^j(n_1, \ldots, n_j) := \frac{1}{G^j} \prod_{i=1}^{j} \prod_{m_i=1}^{n_i} \frac{\lambda_i}{\phi_i^i(m_i)},$$

with $G^j := \sum_{n_1,\ldots,n_j\in\mathbb{N}^j} \prod_{i=1}^{j} \prod_{m_i=1}^{n_i} \frac{\lambda_i}{\phi_i^i(m_i)}$. Defining

$$\tilde{L}_i^{i-1}(\lambda_1, \ldots, \lambda_{i-1}; \phi) := \sum_{n_1,\ldots,n_{i-1}} \ell^{i-1}\phi_i(n_1, \ldots, n_{i-1})\tilde{\pi}^{i-1}(n_1, \ldots, n_{i-1}),$$

we obtain the following simple sufficient stability condition.

**Proposition 3.4** *If $\lambda_j < \tilde{L}_i^{j-1}(\lambda_1, \ldots, \lambda_{i-1}; \phi)$ for all $j = 1, \ldots, l$, then the process $(N_1(t), \ldots, N_l(t))$ is stable.*

**Proof** By virtue of the stochastic dominance and the fact that the allocation function $\phi(\cdot)$ is partially decreasing, we obtain

$$\sum_{n_1,\ldots,n_{i-1}} \ell^{i-1}\phi_i(n_1, \ldots, n_{i-1})\tilde{\pi}^{i-1}(n_1, \ldots, n_{i-1}) \leq L_i^{i-1}(\lambda_1, \ldots, \lambda_{i-1}; \phi).$$

Thus the condition in (3) is implied by the inequality $\lambda_i < \tilde{L}_i^{i-1}(\lambda_1, \ldots, \lambda_{i-1}; \phi)$. □

**Remark 3.2** *Note that the bounds are insensitive to the flow size distribution. Also, the bounds coincide with the exact stability conditions in case of two user classes.*

**Remark 3.3** *Note that the bounds extend to scenarios with several user classes per cell, as long as the correlations between service rates depend on the total number of users per cell only.*

Using (4), the number of users (of class $k$, say) $N_k$ can be bounded by a process $\tilde{N}_k$ having death rates $\ell^k\phi_k(\tilde{Y}^{k-1}, \tilde{N}_k)$, where $\tilde{Y}^{k-1}$ has been defined above and is not influenced by $\tilde{N}_k$. Here we would index the classes in such a manner that class $k$ interferes with classes $1, \ldots, k-1$, but not with classes $k+1, \ldots, K$. Unfortunately, this does not lead directly to closed-form bounds. The difficulty arises from the fact that $\tilde{N}_k$ behaves as a birth-death process driven by a random environment $\tilde{Y}^{k-1}$, which is intractable in general. Hence, we introduce approximations of the bounds along the lines of [14, 32], based on two limit regimes, termed *fluid* and *quasi-stationary*, where the process $\tilde{Y}^{k-1}$ evolves on an infinitely fast and an infinitely slow time scale, respectively. Specifically, we consider a family of systems, parametrized by $s \in (0, \infty)$ and obtained by accelerating the process $\tilde{Y}^{k-1}(t)$ by a factor $s$, i.e., replacing the process $\tilde{Y}^{k-1}(t)$ by $\tilde{Y}^{k-1}(s \times t)$.

**Quasi-stationary regime** The quasi-stationary regime is obtained when the acceleration factor $s$ tends to 0. In the limit for $s \to 0$, the process $\tilde{Y}^{k-1}(t)$ is frozen to its initial state. Thus the quasi-stationary regime corresponds to a scenario where the process $\tilde{Y}^{k-1}(t)$ is constant and equal to $(n_1, \ldots, n_{k-1})$ with probability $\tilde{\pi}^{k-1}(n_1, \ldots, n_{k-1})$.
Assuming $\lambda_k < \inf_{(n_1,\ldots,n_{k-1})\in\mathbb{N}^{k-1}} \liminf_{n_k\to\infty} \ell^k\phi_k(n_1, \ldots, n_{k-1}, n_k)$, we obtain the distribution of the number of active class-$k$ flows in the quasi-stationary regime:

$$p_k^{\mathrm{qs}}(n_k) = \sum_{(n_1,\ldots,n_{k-1})\in\mathbb{N}^{k-1}} \tilde{\pi}^{k-1}(n_1, \ldots, n_{k-1}) p_k^{\mathrm{qs}}(n_k | n_1, \ldots, n_{k-1}),$$

with

$$p_k^{\mathrm{qs}}(n_k | n_1, \ldots, n_{k-1}) = G_k^{\mathrm{qs}}(n_1, \ldots, n_{k-1}) \prod_{m=1}^{n_k} \frac{\lambda_k}{\ell^k\phi_k(n_1, \ldots, n_{k-1}, m)},$$

and $G_k^{\mathrm{qs}}(n_1, \ldots, n_{k-1}) := \left( \sum_{n=0}^{\infty} \prod_{m=1}^{n} \frac{\lambda_k}{\ell^k\phi_k^(n_1,\ldots,n_{k-1},m)} \right)^{-1}$.

**Fluid regime** The fluid regime is obtained when the acceleration factor $s$ tends to $\infty$. In the limit for $s \to \infty$, the process $\tilde{Y}^{k-1}(t)$ evolves so rapidly that when there are $n_k$ class-$k$ users, their total service rate is constant and equal to

$$\bar{\phi}_k^{\mathrm{fl}}(n_k) = \sum_{(n_1,\ldots,n_{k-1})\in\mathbb{N}^{k-1}} \tilde{\pi}^{k-1}(n_1, \ldots, n_{k-1}) \ell^k\phi_k(n_1, \ldots, n_{k-1}, n_k).$$

Assuming $\lambda_k < \liminf_{n_k\to\infty} \bar{\phi}_k^{\mathrm{fl}}(n_k)$, we derive the distribution of the number of active class-$k$ flows in the fluid regime

$$p_k^{\mathrm{fl}}(n_k) = G_k^{\mathrm{fl}} \prod_{m=1}^{n_k} \frac{\lambda_k}{\bar{\phi}_k^{\mathrm{fl}}(m)},$$

with $G_k^{\mathrm{fl}} := \left( \sum_{n=0}^{\infty} \prod_{m=1}^{n} \frac{\lambda_k}{\bar{\phi}_k^{\mathrm{fl}}(m)} \right)^{-1}$.
A question of importance is whether these quasi-stationary and fluid regimes provide actual bounds for the orginal upper bounds. From studies on single-class PS queues with time-varying capacity [32], the performance in the quasi-stationary (resp. fluid) regime is worse (resp. better) than that of the actual system. Numerical results in [7] lend further support to that observation. Thus, the quasi-stationary regime of the upper bound $\tilde{N}^k$ is likely to be an upper bound for the actual number of active flows $N^k$.

# 4 Intra- and inter-cell mobility

So far we have assumed that the rate variations occur on a fast time scale and average out over the time scale of interest for flow-level performance, only manifesting themselves in the throughput gains obtained from channel-aware scheduling. We now turn attention to a scenario where the fluctuations may have a slowly varying component. This component may correspond to the variations in the channel attenuation between BS's and users due to user mobility. In order to simplify the presentation, we ignore fast fading unless otherwise specified. The results presented in this section can be readily extended to account for fast fading. We no longer include time-varying inter-cell interference. Finally, we assume that the scheduling strategies operate on a faster time scale (of the order of ms) than that of user mobility (of the order of several seconds). The results presented in this section offer a summary of those in [8, 21, 22].

## 4.1 Mobility model

In order to model user mobility, it is convenient to adopt a 'state structure', with the users moving among several possible states indexed by a finite set $\mathcal{I}$. The states implicitly correspond to geographic subregions of the network. The latter is divided into a set $\mathcal{B}$ of cells, and we denote by $\mathcal{I}_b$ the set of states corresponding to subregions in cell $b$, i.e., where users are served by BS $b$. When in state $i \in \mathcal{I}_b$, a user is served at rate $C_i$ when scheduled by BS $b$.
We categorize users according to their mobility. Specifically, we consider a system with $K$ traffic classes. We denote by $X_{km}(t)$ the state of the $m$-th arriving class-$k$ user, and by $F_{km}$ its flow size. (For notational convenience, we define $X_{km}(t)$ for all values of $t$. Note however that the $m$-th class-$k$ user may not have arrived yet or may already have departed at time $t$, in which case $X_{km}(t)$ is inconsequential.) We assume that $F_{km}$ and $X_{km}(t)$, $m = 1, 2, \ldots$, are i.i.d. copies of an exponential random variable $F_k$ with mean $\xi_k$ and a Markovian stationary and ergodic process $X_k(t)$ with state space $\mathcal{I}$, respectively. Denote by $R_k(t) = C_{X_k(t)}$ the generic rate process for class $k$. Define $\pi_{i,k} := \mathbb{P}\{X_k(t) = i\}$ as the stationary probability for a class-$k$ user to be in state $i$, and $\mathcal{J}_k := \{i \in \mathcal{I} : \pi_{i,k} > 0\}$ as the set of states in which a class-$k$ user may reside. Define $p_{k,b} := \sum_{i \in \mathcal{I}_b} \pi_{i,k}$ as the stationary probability that a class-$k$ user is in cell $b$. Finally, denote by $\mathcal{B}_k := \{b \in \mathcal{B} : p_{k,b} > 0\}$ the set of cells that class-$k$ users may visit. Users of classes $1, \ldots, K$ generate flows according to independent Poisson processes of intensities $\lambda_1, \ldots, \lambda_K$, respectively, and we denote by $\sigma_k := \lambda_k \xi_k$ the traffic intensity of class-$k$ users.
We first provide stability conditions that depend on the scheduling strategy, and then derive bounds for the transfer delays using stochastic comparison methods.

## 4.2 Stability

We now determine the set of all traffic vectors $(\sigma_1, \ldots, \sigma_K)$ such that the system is stable. This set depends, as we will see, on the underlying scheduling strategy. We first investigate strategies that maximize the stability region, and then consider $\alpha$-fair schedulers.

### 4.2.1 Optimal scheduling strategies

We first determine the capacity region for optimal scheduling strategies. In other words, we characterize the set of all traffic vectors $(\sigma_1, \ldots, \sigma_K)$ for which there exists a scheduling strategy

that achieves stability. Define:

$$\begin{aligned}\mathcal{T} &:= \{(\tau_{i,k})_{i\in\mathcal{I},k=1,\ldots,K} \in \mathbb{R}_+^{(|\mathcal{I}|\times K)} : \tau_{i,k} = 0 \text{ if } i \notin \mathcal{J}_k, \text{ and } \sum_{k=1}^{K} \sum_{i\in\mathcal{I}_b\cap\mathcal{J}_k} \tau_{i,k} = 1, \forall b \in \mathcal{B}\},\\ \mathcal{R} &:= \{(r_1,\ldots,r_K) \in \mathbb{R}_+^K : \exists \tau \in \mathcal{T} \text{ such that } r_k \le \sum_{i\in\mathcal{I}} \tau_{i,k} C_i, \forall k = 1,\ldots,K\}.\end{aligned}$$

The component $\tau_{i,k}$, $i \in \mathcal{I}_b$, may be interpreted as the fraction of resources of BS $b$ allocated to class-$k$ users in state $i$. With that interpretation, the quantity $\sum_{i\in\mathcal{I}} \tau_{i,k} C_i$ represents the total service rate received by class-$k$ users. Thus, $\mathcal{R}$ may be interpreted as the achievable rate region, i.e., the set of all achievable service rates for the various traffic classes.
Note that $\mathcal{R}$ is a convex set and depends on the spatial user distributions through the sets $\mathcal{J}_k$ only.

**Proposition 4.1** *There exists a scheduling strategy that achieves stability if* $(\sigma_1,\ldots,\sigma_K) \in \text{int}(\mathcal{R})$. *Conversely, if* $(\sigma_1,\ldots,\sigma_K) \notin \mathcal{R}$, *then there exists no scheduling strategy that achieves stability.*

**Proof** Assume $(\sigma_1,\ldots,\sigma_K) \in \text{int}(\mathcal{R})$. Then there exists a vector $(\tau_{i,k})_{i,k} \in \mathcal{T}$ such that $\sigma_k < \sum_{i\in\mathcal{I}} \tau_{i,k} C_i$ for all $k = 1,\ldots,K$. Now consider a static scheduling strategy that allocates in cell $b$ a fixed fraction $\tau_{i,k}$, $i \in \mathcal{I}_b$ of the resources to class $k$ in that cell. By considering the system in the fluid limit, it can be shown that this strategy achieves stability. The converse statement follows from the convexity of $\mathcal{R}$. If $(\sigma_1,\ldots,\sigma_K) \notin \mathcal{R}$, then by convexity of $\mathcal{R}$, there exists a linear hyperplane $H$ containing $(\sigma_1,\ldots,\sigma_K)$ defined by $\sum_{k=1}^K a_k\sigma_k = d$, with $a_k, d > 0$, such that there exists an $\epsilon > 0$ with $\sum_{k=1}^K a_k r_k < d - \epsilon$ for all rate vectors $(r_1,\ldots,r_K) \in \mathcal{R}$. Thus, in the fluid limit, the quantity $\sum_{k=1}^K a_k \bar{n_k}(t)$, where $\bar{n_k}(t)$ is the number of active class-$k$ users at time $t$ in the fluid limit, will continuously grow at least at rate $\epsilon$, which implies that the system is unstable. □

**Example 4.1 (Single traffic class)** *If there is just a single traffic class, i.e.,* $K = 1$, *then, dropping the class index* $k$, $\mathcal{R} = \{r \in \mathbb{R}_+ : \exists \tau \in \mathcal{T}$ *such that* $r < \sum_{i\in\mathcal{I}} \tau_i C_i\}$, *which may be represented in a more compact manner as* $\mathcal{R} = [0, r_{\max})$, *with* $r_{\max} := \sum_{b\in\mathcal{B}} \max_{i\in\mathcal{I}_b} C_i$.

**Example 4.2 (Intra-cell mobility only)** *If there is intra-cell mobility only, i.e., each of the sets* $\mathcal{B}_k$ *is just a singleton, then* $\mathcal{R} = \{(r_1,\ldots,r_K) \in \mathbb{R}_+^K : \sum_{k:\mathcal{B}_k=\{b\}} r_k/C_k^{\max} \le 1, \forall b \in \mathcal{B}\}$, *where* $C_k^{\max} := \max_{i\in\mathcal{J}_k} C_i$ *denotes the highest transmission rate of class-*$k$ *users.*

Note that in both the above two examples the BS only serves classes in their most favorable state within the cell.

### 4.2.2 $\alpha$-fair resource sharing

We now assume that each BS implements an $\alpha$-fair scheduling discipline, independently of the behavior of other BS's. Thus when there are $n_{i,k}$ class-$k$ users in state $i$, $i \in \mathcal{I}_b$, each of them receives service at rate $\frac{C_i^{1/\alpha}}{\sum_{l=1}^K \sum_{j\in\mathcal{I}_b} n_{j,l} C_j^{1/\alpha-1}}$.

Define

$$\mathcal{R}^\alpha := \{(r_1,\ldots,r_K)\in\mathbb{R}_+^K : \exists \theta\in\mathbb{R}_+^K \text{ such that } r_k \le \sum_{b\in\mathcal{B}} \frac{\sum_{i\in\mathcal{I}_b}\pi_{i,k}\theta_k C_i^{1/\alpha}}{\sum_{l=1}^K\sum_{j\in\mathcal{I}_b}\pi_{j,l}\theta_l C_j^{1/\alpha-1}}, \forall k=1,\ldots,K\}.$$

Note that the vector $(\tau_{i,k})$ with $\tau_{i,k}=\frac{\pi_{i,k}\theta_k C_i^{1/\alpha-1}}{\sum_{l=1}^K\sum_{j\in\mathcal{I}_b}\pi_{j,l}\theta_l C_j^{1/\alpha-1}}$, $i\in\mathcal{I}_b$, belongs to the set $\mathcal{T}$, so that $\mathcal{R}^\alpha\subseteq\mathcal{R}$.
The components of the vector $(\theta_1,\ldots,\theta_K)$ may be interpreted as the numbers of flows of the various classes. With that interpretation, the quantity $\sum_{b\in\mathcal{B}}\frac{\sum_{i\in\mathcal{I}_b}\pi_{i,k}\theta_k C_i^{1/\alpha}}{\sum_{l=1}^K\sum_{j\in\mathcal{I}_b}\pi_{j,l}\theta_l C_j^{1/\alpha-1}}$ represents the total service rate received by class-$k$ flows under an $\alpha$-fair sharing strategy.
The region $\mathcal{R}^\alpha$ has a non-linear boundary in general, and may either be convex or non-convex, depending on the values of the probabilities $\pi_{i,k}$ and the rates $C_i$, see Example 4.5 for a graphical illustration.
The next proposition provides a characterization of the capacity region in case of $\alpha$-fair resource sharing, assuming exponential service requirements. It states that the system is stable if and only if there exists a relative distribution of the numbers of flows across the various classes so that the total service rate received by each of the traffic classes is larger than the traffic intensity of that class.

**Proposition 4.2** *If $(\sigma_1,\ldots,\sigma_K)\in\text{int}(\mathcal{R}^\alpha)$, then the system is stable. If $(\sigma_1,\ldots,\sigma_K)\notin\mathcal{R}^\alpha$, then the system is unstable.*

**Proof** The proof relies on the consideration of fluid limits [31], where systems with a large population of flows are considered. In such limiting systems, it can be shown that the total service rate received by class-$k$ flows is given by:

$$r_k(t)=\sum_{b\in\mathcal{B}}\frac{\bar{n}_k(t)A_{k,b}^\alpha}{\sum_{l=1}^K\bar{n}_l(t)B_{l,b}^\alpha}, \tag{5}$$

for $\bar{n}_k(t)>0$, and where

$$A_{k,b}^\alpha=\sum_{i\in\mathcal{I}_b}\pi_{i,k}C_i^{1/\alpha},\quad B_{k,b}^\alpha=\sum_{i\in\mathcal{I}_b}\pi_{i,k}C_i^{1/\alpha-1}.$$

In the above formula, $\bar{n}_k(t)$ denotes the number of class-$k$ flows at time $t$. The proof of (5) is standard when a single cell is considered, but involves a spatial homogeneity property in the case of networks with several cells. Intuitively, one can justify this formula by observing that when the number of class-$k$ flows is very large then at any instant, the number of such flows in state $i$ should be $\pi_{i,k}\bar{n}_k(t)$. This statement can be formally justified as in [56].
The evolution of the fluid limit is characterized by the following set of differential equations:

$$\forall k,\quad \frac{d}{dt}\bar{n}_k(t)=\lambda_k-\mu_k\times r_k(t),$$

where $\mu_k:=1/\xi_k$.
Sufficient stability condition

Assume that $(\sigma_1, \ldots, \sigma_K) \in \text{int}(\mathcal{R}^\alpha)$. Then there exists a vector $(\theta_1, \ldots, \theta_K) \in \mathbb{R}_+^K$ such that $\lambda_k < \mu_k \sum_{b \in \mathcal{B}} \frac{\theta_k A^\alpha_{k,b}}{\sum_{l=1}^K \theta_l B^\alpha_{l,b}} - \epsilon$ for $\epsilon > 0$ sufficiently small for all $k = 1, \ldots, K$. We now consider the quantity $y(t) := \max_{k=1,\ldots,K} \bar{n}_k(t)/\theta_k$ and show that it will continuously decrease at a strictly negative rate.

Denote $k^* := \arg\max_{k=1,\ldots,K} \bar{n}_k(t)/\theta_k$, where the implicit dependence on $t$ is suppressed. Then:

$$\begin{aligned}
\frac{d}{dt}\bar{n}_{k^*}(t) &= \lambda_{k^*} - \mu_{k^*} \sum_{b \in \mathcal{B}} \frac{\bar{n}_{k^*}(t) A^\alpha_{k^*,b}}{\sum_{l=1}^K \bar{n}_l(t) B^\alpha_{l,b}} \\
&= \lambda_{k^*} - \mu_{k^*} \sum_{b \in \mathcal{B}} \frac{A^\alpha_{k^*,b}}{\sum_{l=1}^K B^\alpha_{l,b} \bar{n}_l(t)/\bar{n}_{k^*}(t)} \\
&< \lambda_{k^*} - \mu_{k^*} \sum_{b \in \mathcal{B}} \frac{A^\alpha_{k^*,b}}{\sum_{l=1}^K B^\alpha_{l,b} \theta_l/\theta_{k^*}} \\
&= \lambda_{k^*} - \mu_{k^*} \sum_{b \in \mathcal{B}} \frac{A^\alpha_{k^*,b} \theta_{k^*}}{\sum_{l=1}^K B^\alpha_{l,b} \theta_l} \\
&< \epsilon,
\end{aligned}$$

whenever $\bar{n}_{k^*}(t) > 0$. We conclude that the fluid limit reaches zero in finite time, and hence the system is stable [31].

Necessary stability condition

Denote by $\partial\mathcal{R}^\alpha$ the boundary of $\mathcal{R}^\alpha$, i.e., $r \in \partial\mathcal{R}^\alpha$ if $r \in \mathcal{R}^\alpha$ and there exists a $k$ such that $\forall \epsilon > 0$, $r + \epsilon.e_k \notin \mathcal{R}^\alpha$. This boundary is the union of the surfaces $\partial\mathcal{R}^\alpha_{\mathcal{L}}$ over all non-empty subsets $\mathcal{L}$ of $\{1, \ldots, K\}$, with:

$$\begin{aligned}
\partial\mathcal{R}^\alpha_{\mathcal{L}} = \Bigg\{ & r \in \mathbb{R}_+^K : \exists \theta \in \mathbb{R}_+^{|\mathcal{L}|}, \exists \theta' \in \mathbb{R}_+^{|\bar{\mathcal{L}}|}, \\
& \forall k \in \mathcal{L}, \theta_k > 0, r_k = \sum_{b \in \mathcal{B}} \frac{\theta_k A^\alpha_{k,b}}{\sum_{l \in \mathcal{L}} \theta_l B^\alpha_{l,b}}, \\
& \forall k \in \bar{\mathcal{L}}, r_k \leq \sum_{b \in \mathcal{B}^{\mathcal{L}}} \frac{\theta'_k A^\alpha_{k,b}}{\sum_{l \in \bar{\mathcal{L}}} \theta'_l B^\alpha_{l,b}} \Bigg\},
\end{aligned}$$

where $\bar{\mathcal{L}} = \{1, \ldots, K\} \setminus \mathcal{L}$ and $\mathcal{B}^{\mathcal{L}}$ denotes the set of cells that are not visited by users of classes in the set $\mathcal{L}$. The surface $\partial\mathcal{R}^\alpha_{\mathcal{L}}$ is the set of points of $\mathcal{R}^\alpha$ parametrized by $\theta$ such that $\forall k \in \mathcal{L}$, $\forall l \in \bar{\mathcal{L}}$, $\theta_k \gg \theta_l$. It is worth remarking that $\partial\mathcal{R}^\alpha_{\mathcal{L}}$ is part of a cylinder with directions parallel to the components in $\bar{\mathcal{L}}$, i.e., if $r$ is a point of this surface, then $r' = (r_k, k \in \mathcal{L},\ r'_l,\ l \in \bar{\mathcal{L}})$ is also a point of this surface provided that for all $l \in \bar{\mathcal{L}}$, $0 \leq r'_l \leq r_l$. It is also important to note that in the definition of $\partial\mathcal{R}^\alpha_{\mathcal{L}}$, we can choose $\theta \in \mathbb{R}_+^{|\mathcal{L}|}$ with strictly positive components; this is because the points obtained when some of the components of this $\theta$ are equal to 0, are included in some other surfaces with a different set $\mathcal{L}$.

We now prove that the system is unstable by induction on the number $K$ of flow classes. The result holds for $K = 1$. Assume that it is true for all systems with at most $K - 1$ classes. Let us prove it in the case of $K$-class systems. Assume that $\sigma = (\sigma_1, \ldots, \sigma_K) \notin \mathcal{R}^\alpha$. Without loss of generality, we can assume that $\sigma_k > 0$ for all $k = 1, \ldots, K$. Now define $\gamma$ as the maximum real number such that $\gamma \times \sigma \in \mathcal{R}^\alpha$. By assumption, $\gamma < 1$. Of course, we have $\gamma \times \sigma \in \partial\mathcal{R}^\alpha$, and there exists a set of classes $\mathcal{L}$ such that $\gamma \times \sigma \in \partial\mathcal{R}^\alpha_{\mathcal{L}}$. We consider two cases:

(i) If $\mathcal{L} = \{1, \ldots, K\}$, then we deduce that there exists $\theta \in \mathbb{R}_+^K$ such that $\theta_k > 0$ for all $k = 1, \ldots, K$ and:

$$\sigma_k > \sum_{b \in \mathcal{B}} \frac{\theta_k A_{k,b}^\alpha}{\sum_{l=1}^K \theta_l B_{l,b}^\alpha}.$$

We deduce that the system is unstable. Indeed one can easily show that the fluid limit grows at least linearly to $\infty$.
(ii) Otherwise, we consider the restricted system where the classes in $\bar{\mathcal{L}}$ have no traffic. Note that the restricted system provides a stochastic lower bound of the actual system (this is due to the fact that all the systems considered are monotonic [14]). Hence we just need to prove that the restricted system is unstable. Note that the projection of $\partial\mathcal{R}_{\mathcal{L}}^\alpha$ on the sub-space of $\mathcal{L}$ components is actually the boundary of the set $\mathcal{R}_{\mathcal{L}}^\alpha$ defined by:

$$\mathcal{R}_{\mathcal{L}}^\alpha = \left\{ r \in \mathbb{R}_+^{|\mathcal{L}|} : \exists \theta \in \mathbb{R}_+^{|\mathcal{L}|}, \forall k \in \mathcal{L}, r_k \leq \sum_{b \in \mathcal{B}} \frac{\theta_k A_{k,b}^\alpha}{\sum_{l \in \mathcal{L}} \theta_l B_{l,b}^\alpha} \right\}.$$

Since $\partial\mathcal{R}_{\mathcal{L}}^\alpha$ is a cylinder, we deduce that $(\sigma_k, k \in \mathcal{L}) \notin \mathcal{R}_{\mathcal{L}}^\alpha$, and finally that the restricted system is unstable by induction. □

**Example 4.3 (Single traffic class)** *If there is just a single traffic class, i.e., $K = 1$, then $\mathcal{R}^\alpha = [0, r_{\max}^\alpha)$ with, dropping the class index $k$, $r_{\max}^\alpha := \sum_{b \in \mathcal{B}} C_b$, with $C_b := \sum_{i \in \mathcal{I}_b} \pi_i C_i^{1/\alpha} / \sum_{j \in \mathcal{I}_b} \pi_j C_j^{1/\alpha - 1}$.*

**Example 4.4 (Intra-cell mobility only)** *If there is intra-cell mobility only, i.e., each of the sets $\mathcal{B}_k$ is just a singleton, then $\mathcal{R}^\alpha = \{(r_1, \ldots, r_K) \in \mathbb{R}_+^K : \sum_{k:\mathcal{B}_k = \{b\}} r_k / C_k^{\text{ave}} \leq 1, \forall b \in \mathcal{B}\}$, where $C_k^{\text{ave}} := \sum_{i \in \mathcal{J}_k} \pi_{i,k} C_i$ denotes the average rate coefficient of class-$k$ flows.*

In both the above two examples it is easily seen that the capacity region $\mathcal{R}^\alpha$ *decreases* when the value of the fairness index $\alpha$ *increases*, which in fact holds in greater generality as will be shown in Proposition 4.3 below.

**Example 4.5 (A 3-cell 3-class network)** *In general, the capacity region $\mathcal{R}^\alpha$ has non-linear boundaries and its shape turns out to be rather intricate. We present an example of a three-cell three-class network to illustrate how astonishing this shape can be. There are five states, with $C_1 = C_4 = C_5 = 2$, $C_2 = C_3 = 1$, $\mathcal{I}_1 = \{1, 2\}$, $\mathcal{I}_2 = \{3, 4\}$, and $\mathcal{I}_3 = \{5\}$. Class-1 users oscillate between states 1 and 3 with probability 0.1 to be in state 1, class-2 users between states 2 and 4 with probability 0.9 to be in state 2, and class-3 users between states 2 and 5, with equal probability to be in each state. Figure 2 depicts $\mathcal{R}^1$, i.e., the capacity region for the PF strategy, and indicates the various surfaces $\partial\mathcal{R}_1^{\mathcal{L}}$ composing the boundary of $\mathcal{R}^1$.*

### 4.2.3 Impact of scheduling

It is worth observing that the stability region $\mathcal{R}^\alpha$ depends on the fairness index $\alpha$, i.e., on the scheduling discipline, even in the simple case of single-cell single class systems. This contrasts with the case where users are not moving (see § 4.2.4). The next proposition, whose proof can be found in [22], states that the stability region actually grows monotonically as the value of $\alpha$ decreases.

**Proposition 4.3** *If $\alpha' \geq \alpha$ then $\mathcal{R}^{\alpha'} \subseteq \mathcal{R}^\alpha$.*

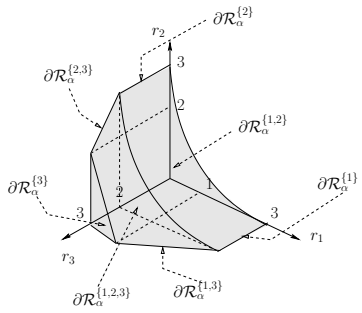



Figure 2: Capacity region $\mathcal{R}^1$ of a three-cell three-class network with mobility under the PF strategy.

### 4.2.4 Mobility increases stability

Let us first provide the stability region in the case where users are not moving, i.e., the state of a user is assumed to be fixed for the duration of the flow, and the probability that a class-$k$ user is in state $i$ is $\psi_{i,k}$. Consider systems where the scheduling discipline implemented at each BS is arbitrary but work-conserving, i.e., each BS is active whenever there is a user in the corresponding cell. It can be easily shown that the stability region without mobility does not depend on the scheduling discipline and simplifies to

$$\mathcal{R}^{\text{no}} := \{(r_1, \ldots, r_K) \in \mathbb{R}_+^K : \exists \tau \in \mathcal{T} \text{ such that } \psi_{i,k} r_k < \tau_{i,k} C_i, \forall i \in \mathcal{I}, k = 1, \ldots, K\},$$

which may be represented in a more compact manner as

$$\mathcal{R}^{\text{no}} = \{(r_1, \ldots, r_K) \in \mathbb{R}_+^K : \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_b} \psi_{i,k} r_k / C_i < 1, \forall b \in \mathcal{B}\}.$$

To compare the stability regions with and without mobility, it makes sense to assume that for all $k$, $i$, $\pi_{i,k} = \psi_{i,k}$. The next proposition first states that the stability region with mobility is larger than that without mobility as long as the scheduling discipline is left flexible, which agrees with findings in the context of ad-hoc mobile networks [35]. It further states that this observation also remains valid for $\alpha$-fair scheduling strategies. Its proof can be found in [22].

**Proposition 4.4** *We have:*

$$\mathcal{R}^{\text{no}} \subseteq \mathcal{R}, \quad \textit{and for any} \quad \alpha > 0, \ \mathcal{R}^{\text{no}} \subseteq \mathcal{R}^\alpha.$$

## 4.3 Transfer delays

With user mobility, the network may be interpreted as a system of (PS) queues with state-dependent and time-varying capacities, and not surprisingly, it proves impossible to obtain exact expressions for the transfer delays (even the average values). In this subsection, we develop a method, based on stochastic comparison techniques, to derive upper and lower bounds for transfer delays. We restrict our attention to intra-cell mobility, and consider the class of schedulers such that the service rate of a class-$k$ user in state $i$ is of the form:

$$C_i \times H_k(n_1, \ldots, n_K),$$

where as before $n_k$ denotes the number of active class-$k$ users. For any $k$, the function $H_k(\cdot)$ is decreasing in each of the $n_j$'s, $j = 1, \ldots, K$. Note that the above form does not apply for $\alpha$-fair schedulers, except the PF strategy ($\alpha = 1$), for which $H_k(n_1, \ldots, n_K) = G(\sum_{k=1}^K n_k)/(\sum_{k=1}^K n_k)$. For general $\alpha$-fair schedulers, we believe that the results remain valid, but the proof will require a different approach.

### 4.3.1 Limiting regimes

In order to obtain tractable performance estimates, we introduce two limit regimes, termed *fluid* 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, parametrized by $s \in (0, \infty)$, where the generic rate process for class $k$ is $R_k^{(s)}(t) := R_k(st)$. 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 the case where $X_k(t)$ 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 $R_k^{\rm fl}(t) := R_k^{(\infty)}(t) = \bar{R}_k$, where $\bar{R}_k := \mathbb{E}\{R_k(0)\}$. It is worth observing that the fluid regime is reminiscent of (but different from) the usual law-of-large-numbers fluid limit. 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 $R_k^{\rm qs}(t) := R_k^{(0)}(t) = R_k(0)$, where $R_k(0)$ has the stationary marginal distribution of the process $R_k(t)$.

For example, for the PF strategy, the fluid and quasi-stationary regimes yield tractable expressions for the transfer delays. Define the traffic intensities associated with class $k$ in the fluid and quasi-stationary regimes as $\rho_k^{\rm fl} := \lambda_k \xi_k / \bar{R}_k$ and $\rho_k^{\rm qs} := \lambda_k \xi_k \mathbb{E}\{1/R_k(0)\}$, respectively. Note that these values 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 $\rho^{\rm qs} := \sum_{k=1}^K \rho_k^{\rm qs}$ the total traffic intensities in the fluid and quasi-stationary regimes, respectively. With the above translation to a multi-class PS system (without time-varying capacities), the performance in the fluid and quasi-stationary regimes may be explicitly evaluated using results of Subsection 2.1. In particular, 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 respective regimes depend on the class characteristics through the traffic intensities $\rho_k^{\rm fl}$ and $\rho_k^{\rm qs}$, respectively, 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.

### 4.3.2 Stochastic comparisons

We now prove that under rather mild assumptions, the fluid and quasi-stationary regimes yield lower and upper bounds for the performance of the system with user mobility. The reason why these two regimes provide bounds is because - informally speaking - the adverse impact of getting stuck in a bad state for a long time far outweighs the beneficial effect of residing in a good state for a long period, and rapid averaging is the most favorable situation. Under additional assumptions, we actually show the stronger property that the performance improves when the speed parameter $s$ is increased.
Define the total workload of the system at time $t$ as:

$$W(t) = \sum_{k=1}^{K} \sum_{j=1}^{N_k(t)} \frac{F_{kj}(t)}{\bar{R}_k},$$

where $N_k(t)$ denotes the number of active class-$k$ users at time $t$, and $F_{kj}(t)$ is the remaining size of active class-$k$ flow $j$ at time $t$. We also denote by $W$, $N_k$, and $S_k$ the workload, the number of class-$k$ flows, and the class-$k$ transfer delay in equilibrium (with some abuse of notation, we assume that these quantities are infinite when the system is unstable). Similarly, we introduce $W^{\mathrm{fl}}$, $W^{\mathrm{qs}}$, $W^{(s)}$, $N_k^{\mathrm{fl}}$, $N_k^{\mathrm{qs}}$, $N_k^{(s)}$, $S_k^{\mathrm{fl}}$, $S_k^{\mathrm{qs}}$, $S_k^{(s)}$.
*Assumption 4.1:* The cumulative distribution function $P(\cdot) = \mathbb{P}\{F_k \leq \cdot\}$ of class-$k$ flow sizes is concave.
*Assumption 4.2:* The rate process of class-$k$ users is a homogeneous stationary Markov process. The transition kernels $Q$ and $Q_r$ of the rate process and of its time-reversed process are *st*-monotone in the following sense: for all increasing functions $f(\cdot)$, $x \mapsto \int f(u)Q(x, du)$ is also increasing.

**Proposition 4.5** *Under Assumption 4.1, we have:*

$$W^{\mathrm{fl}} \leq_{icx} W \leq_{icx} W^{\mathrm{qs}},$$

*and for all* $k = 1, \ldots, K$,

$$N_k^{\mathrm{fl}} \leq_{st} N_k \leq_{st} N_k^{\mathrm{qs}},$$
$$S_k^{\mathrm{fl}} \leq_{st} S_k \leq_{st} S_k^{\mathrm{qs}}.$$

**Proposition 4.6** *Under Assumptions 4.1 and 4.2, we have: for any* $s > 1$,

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

*and for all* $k = 1, \ldots, K$,

$$N_k^{(s)} \leq_{st} N_k, \quad S_k^{(s)} \leq_{st} S_k.$$

In the above propositions, $\leq_{st}$ and $\leq_{icx}$ denote the strong and *icx* stochastic orders (for two r.v. $X$ and $Y$, we have $X \leq_{icx} Y$ if and only if for all increasing and convex functions $f(\cdot)$ such that $\mathbb{E}\{f(Y)\} < \infty$, $\mathbb{E}\{f(X)\} \leq \mathbb{E}\{f(Y)\}$). In order to illustrate the methods used to prove the above propositions, we next provide the proof of Proposition 4.5 (the proof of Proposition 4.6 is given in [8]).

### 4.3.3 Proof of Proposition 4.5

We prove the inequalities of Proposition 4.5 for the workload $W(t)$, the number of class-$k$ flows $N_k(t)$ at time $t$, and the transfer delay $S$ of an arbitrary flow, in a system initially empty. The results in equilibrium are deduced using classical results about the stability of the strong stochastic order by limits, and for the workload, using the monotonicity property of Loynes' construction (the *icx* order is not stable by limits).
We first prove the inequalities in the following slotted system. The interval is divided into 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 inequalities for a non-slotted system then follow 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 $\infty$.
The proof is based on the notion of supermodular functions (see, e.g., [53]) and on the Lorentz inequality [62]. To simplify the notation, we assume that $H_k$ depends only on the total number of active flows $n$, and that $H_k(n) = G(n)/n$. All proofs remain unchanged without this convenient assumption.

**Definition 4.1** *(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 4.1** *(Lorentz inequality) Let* $Z_1, \ldots, Z_n$ *be identically distributed r.v.'s. For all supermodular functions* $f(\cdot)$, $\mathbb{E}\{f(Z_1, \ldots, Z_n)\} \leq \mathbb{E}\{f(Z_1, \ldots, Z_1)\}$.

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 $\mathbb{E}_1$ 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 4.2** *For all* $k = 1, \ldots, K$, $\mathbb{E}_1[f(W_L)]$ *and* $\mathbb{E}_1[g(N_{k,L})]$ *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 $(x_l, \ldots, x_L)$. 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 $F_1 \leq x_l G(n_l)/n_l$, $W_L = w_l$ and $N_{k,L} = n_{k,l}$;
- for $l' = l+1, \ldots, L-1$, if $\sum_{j=l}^{l'-1} x_j G(n_j)/n_j < F_1 \leq \sum_{j=l}^{l'} x_j G(n_j)/n_j$, $W_L = w_{l'}$ and $N_{k,L} = n_{k,l'}$;
- if $F_1 > \sum_{j=l}^{L-1} x_j G(n_j)/n_j$, $W_L = w_L + W_L^1$ and $N_{k,L} = n_{k,L}$.

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

$$\mathbb{E}_1\left[f(W_L)\right] = (f(w_l) - f(w_{l+1}))\mathbb{P}_1\left[x_l \frac{G(n_l)}{n_l}\right]$$

$$+ \cdots + (f(w_{L-1}) - f(w_L))\mathbb{P}_1\left[\sum_{j=l}^{L-1} x_j \frac{G(n_j)}{n_j}\right] + f(w_L)\mathbb{P}_1\left[\sum_{j=l}^{L-1} x_j \frac{G(n_j)}{n_j}\right]$$

$$+ \mathbb{E}_1\left[f(w_L + W_L^1)\mathrm{I}_{\{F_1 > \sum_{j=l}^{L-1} x_j \frac{G(n_j)}{n_j}\}}\right],$$

where $\mathbb{P}_1$ 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 $\mathbb{E}_1\left[f(w_L + W_l^1)\right]$, which is a supermodular and convex function of $(x_l, \ldots, x_L)$. In addition, it follows from Assumption 4.1 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. As $G(n)/n$ decreases in $n$, we have $w_l \leq \cdots \leq w_L$, so that $\mathbb{E}_1\left[f(W_L)\right]$, as the sum of supermodular and convex functions, is supermodular and convex.
Similarly, we have for all increasing functions $g(\cdot)$:

$$\mathbb{E}_1\left[g(N_{k,L})\right] = (g(n_{k,l}) - g(n_{k,l+1}))\mathbb{P}_1\left[x_l \frac{G(n_l)}{n_l}\right]$$

$$+ \cdots + (g(n_{k,L-1}) - g(n_{k,L}))\mathbb{P}_1\left[\sum_{j=1}^{L-1} x_j \frac{G(n_j)}{n_j}\right] + g(n_{k,L}).$$

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

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

$$\mathbb{E}\{f(W_L^{\mathrm{fl1}})\} = \mathbb{E}\{\mathbb{E}_1\left[f(W_L)\right](\mathbb{E}\{x_l\}, \ldots, \mathbb{E}\{x_l\})\},$$

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

$$\mathbb{E}\{f(W_L^{\mathrm{qs1}})\} = \mathbb{E}\{\mathbb{E}_1\left[f(W_L)\right](x_l, \ldots, x_l)\},$$

where $^{\mathrm{fl1}}$ (resp. $^{\mathrm{qs1}}$) denotes the fluid (resp. 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 4.2 and the fact that $(\mathbb{E}\{x_l\}, \ldots, \mathbb{E}\{x_l\}) \leq_{icx} (x_l, \ldots, x_L)$

that for all increasing and convex functions $f(\cdot)$, all increasing functions $g(\cdot)$, and all $k = 1, \ldots, K$:

$$\mathbb{E}\{f(W_L^{\mathrm{fl1}})\} \leq \mathbb{E}\{f(W_L)\},\ \mathbb{E}\{g(N_{k,L}^{\mathrm{fl1}})\} \leq \mathbb{E}\{g(N_{k,L})\},$$

Similarly, using the independence of the rate processes, it follows from Lemma 4.2 and the Lorentz inequality that:

$$\mathbb{E}\{f(W_L)\} \leq \mathbb{E}\{f(W_L^{\mathrm{qs1}})\},\ \mathbb{E}\{g(N_{k,L})\} \leq \mathbb{E}\{g(N_{k,L}^{\mathrm{qs1}})\}.$$

We obtain the required inequalities for the workload and number of flows by applying successively the same reasoning to an arbitrary set of flows satisfying Assumption 4.1, i.e., to all flows here.

We now prove the inequalities regarding the transfer delay. Let $S_L$ be the time spent by an arbitrary flow in the slotted system up to slot $L$. We prove exactly as in Lemma 4.2 that that for all increasing functions $g(\cdot)$, $\mathbb{E}_1[g(S_L)]$ is a supermodular and convex function of $(x_1, \ldots, x_L)$. We deduce as above that:

$$\mathbb{E}\{g(S_L^{\mathrm{fl}})\} \leq \mathbb{E}\{g(S_L)\} \leq \mathbb{E}\{g(S_L^{\mathrm{qs}})\},$$

and by letting $L$ tend to $\infty$,

$$\mathbb{E}\{g(S(t)^{\mathrm{fl}})\} \leq \mathbb{E}\{g(S(t))\} \leq \mathbb{E}\{g(S(t)^{\mathrm{qs}})\},$$

where $S(t)$ denotes the time spent by an arbitrary flow in the non-slotted system up to time $t$. We obtain the required inequalities by letting $t$ tend to $\infty$.

### 4.3.4 Numerical experiments

To conclude this section, we illustrate the results derived above in the case of a simple multi-cell linear network. In this network, which may for example represent a road, BS's are regularly spaced, and users move at the same speed in both directions (there are two flows classes). Flow instances, or file transfers, are generated at positions uniformly distributed in the network, with mean flow sizes of 75 kbytes. The traffic intensities of the two flow classes are identical. The feasible rate of a flow can take three different values, 4 Mb/s close to BS's, 3 Mb/s, and 2 Mb/s at the cell edges. Cells are of diameter equal to 1 km. This setting roughly corresponds to CDMA 1xEV-DO systems.

For this network, we calculate the stability condition without user mobility to be a traffic intensity less than 2.74 Mb/s per cell. In order to assess the impact of the fairness index $\alpha$ and of mobility on the transfer delay, we use the notion of mean flow throughput defined as the ratio of the mean flow size to the mean transfer delay. Figure 3 (left) presents the mean flow throughput as a function of the total traffic intensity for a user speed equal to 90 km/h, and for various values of the fairness index $\alpha$. The performance improves when the value of $\alpha$ is decreased, as does the stability limit. The latter can be significantly increased by choosing a scheduler with $\alpha$ close to 0. In Figure 3 (right), we show that for a fixed scheduler ($\alpha$ is close to 0), increasing the speed of users improves the performance. This suggests that the stochastic comparison results derived in Subsection 4.3.2 also hold in the case of networks with inter-cell mobility.

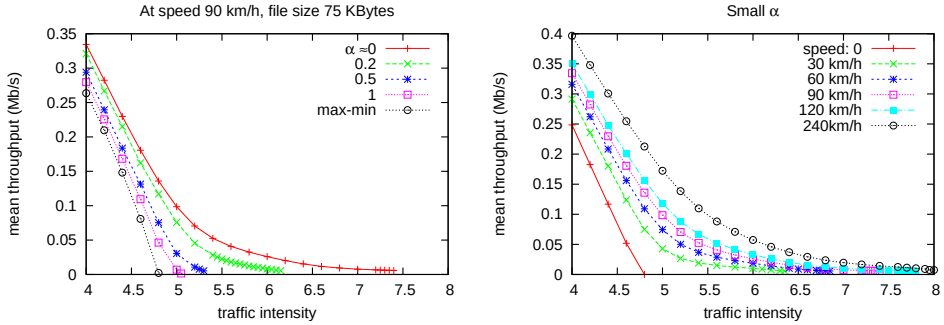



Figure 3: Mean throughput in a linear network- Left: mean user speed of 90 km/h, varying $\alpha$ - Right: small $\alpha$, varying speed.

# 5 Conclusion

We have reviewed several models for characterizing the capacity and evaluating the flow-level performance of wireless networks carrying elastic data transfers. We established that a wide family of so-called $\alpha$-fair channel-aware scheduling strategies provide maximum stability, and described how the special case of the Proportional Fair policy gives rise to a Processor-Sharing model with a state-dependent service rate. We further derived both necessary and sufficient stability conditions for a network of several base stations with inter-cell interference, and constructed lower and upper bounds for the flow-level performance measures. Lastly we investigated the impact of user mobility that occurs on a slow time scale and causes possible hand-offs of active sessions. It was shown that the mobility tends to increase the capacity region, both in the case of globally optimal scheduling and local $\alpha$-fair scheduling. We additionally demonstrated that the capacity and user throughput improve with lower values of the fairness index $\alpha$.

Several potential avenues for further research present themselves. First of all, it would be interesting to consider networks with both inter-cell interference and user mobility, which were treated separately in the present paper. A further appealing problem would be to examine whether intra-cell mobility improves the performance for general $\alpha$-fair strategies as well. A final important challenge would be to prove that inter-cell mobility improves not only the capacity but also the performance in terms of transfer delays and user throughputs as we observed in the numerical experiments.

**Acknowledgment** The authors are grateful to Jim Roberts for his encouragement, support and insightful observations and suggestions.

# References

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

[2] D.M. Andrews (2004). Instability of the Proportional Fair scheduling algorithm for HDR. *IEEE Trans. Wireless Commun.* **3**, 1422–1426.

[3] D.M. Andrews, K. Kumaran, K. Ramanan, A.L. Stolyar, R. Vijayakumar, P.A. Whiting (2004). Scheduling in a queueing system with asynchronously varying service rates. *Prob. Eng. Inf. Sc.* **18**, 191–217.

[4] D.M. Andrews, L. Qian, A.L. Stolyar (2005). Optimal utility-based throughput allocation subject to throughput constraints. In: *Proc. Infocom 2005*.

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

[6] T. Bonald (2004). A score-based opportunistic scheduler for fading radio channels. In: *Proc. European Wireless Conf. 2004.*

[7] T. Bonald, S.C. Borst, N. Hegde, A. Proutière (2004). Wireless data performance in multi-cell scenarios. In: *Proc. ACM Sigmetrics / Performance 2004*, 378–388.

[8] T. Bonald, S.C. Borst, A. Proutière (2004). How mobility impacts the flow-level performance of wireless data systems. In: *Proc. IEEE Infocom 2004*.

[9] T. Bonald, S.C. Borst, A. Proutière (2005). Inter-cell scheduling in wireless data networks. In: *Proc. European Wireless Conf. 2005*.

[10] T. Bonald, S.C. Borst, A. Proutière (2006). Inter-cell coordination in wireless data networks. In: *European Trans. Telecommun.* **17**, 303–312.

[11] T. Bonald, L. Massoulié (2001). Impact of fairness on Internet performance. In: *Proc. ACM Sigmetrics / Performance 2001*, 82–91.

[12] T. Bonald, L. Massoulié, A. Proutière, J. Virtamo (2003). A queueing analysis of max-min fairness, proportional fairness and balanced fairness. *Queueing Systems* **53**, 65–84.

[13] T. Bonald, A. Proutière (2003). Wireless downlink data channels: User performance and cell dimensioning. In: *Proc. ACM Mobicom 2003*, 339–352.

[14] T. Bonald, A. Proutière (2004). On performance bounds for the integration of elastic and adaptive streaming flows. In: *Proc. ACM Sigmetrics / Performance 2004*, 235–245.

[15] T. Bonald, A. Proutière (2006). Flow-level stability of utility-based allocations for non-convex rate regions. In: *Proc. CISS 2006*.

[16] T. Bonald, A. Proutière, G. Régnié, J.W. Roberts (2001). Insensitivity results in statistical bandwidth sharing. 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), 125–136.

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

[18] S.C. Borst (2005). User-level performance of channel-aware scheduling algorithms in wireless data networks. *IEEE/ACM Trans. Netw.* **13**, 636–647.

[19] S.C. Borst (2008). Flow-level performance and user mobility in wireless data networks. *Phil. Trans. Roy. Soc. A* **366 (1872)**, 2047–2058.

[20] S.C. Borst, N. Hegde (2007). Integration of streaming and elastic traffic in wireless networks. In: *Proc. Infocom 2007*.

[21] S.C. Borst, N. Hegde, A. Proutière (2006). Capacity of wireless networks with intra- and inter-cell mobility. In: *Proc. Infocom 2006*.

[22] S.C. Borst, N. Hegde, A. Proutière (2009). Mobility-driven scheduling in wireless networks. In: *Proc. Infocom 2009*, to appear.

[23] S.C. Borst, M. Jonckheere (2006). Flow-level stability of channel-aware scheduling-algorithms. In: *Proc. WiOpt '06*.

[24] S.C. Borst, M. Jonckheere, L. Leskelä (2008). Stability of parallel queueing systems with coupled service rates. *Disc. Event Dyn. Syst.* **18**, 447–472.

[25] S.C. Borst, P.A. Whiting (2001). Dynamic rate control algorithms for HDR throughput optimization. In: *Proc. Infocom 2001*, 976–985.

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

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

[28] J.W. Cohen, O.J. Boxma (1983). *Boundary Value Problems in Queueing System Analysis.* (North-Holland Publ. Cy., Amsterdam).

[29] J.W. Cohen (1984). On a functional relation in three complex variables; three coupled processors. Technical Report 359, Mathematical Institute, University of Utrecht.

[30] J.W. Cohen (1979). The multiple phase service network with generalized processor sharing. *Acta Informatica* **12**, 245–284.

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

[32] F. Delcoigne, A. Proutière, G. Régnié (2004). Modelling integration of streaming and data traffic. *Perf. Eval.* **55**, 185–209.

[33] G. Fayolle, R. Iasnogorodski (1979). Two coupled processors: the reduction to a Riemann-Hilbert problem. *Z. Wahr. verw. Geb.* **47**, 325–351.

[34] L. Georgiadis, M.J. Neely, L. Tassiulas (2006). Resource allocation and cross-layer control in wireless networks. *Found. Trends Netw.* **1**, 1–144.

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

[36] J. Hansen, C. Reynolds, S. Zachary (2007). Stability of processor sharing networks with simultaneous resource requirements. *J. Appl. Prob.* **44**, 636–651.

[37] A. Jalali, R. Padovani, R. Pankaj (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.

[38] M. Jonckheere, S.C. Borst (2006) Stability of multi-class queueing systems with state-dependent service rates. In: *Proc. ValueTools 2006*, Pisa, Italy, October 11–13.

[39] M. Jonckheere (2008). Stability of two interfering processors with load balancing. In: *Proc. ValueTools 2008*. Athens, Greece, October 20–24.

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

[41] F.P. Kelly, A. Maulloo, D. Tan (1998). Rate control for communication networks: Shadow prices, proportional fairness and stability. *J. Oper. Res. Soc.* **49**, 237–252.

[42] P.B. Key, L. Massoulié, A. Bain, F.P. Kelly (2004). Fair Internet traffic integration: network flow models and analysis. *Ann. Telecommun.* **59**, 1338–1352.

[43] R. Knopp, P.A Humblet (1995). Information theory and power control in single-cell multi-user communications. In: *Proc. ICC '95*.

[44] H.J. Kushner, P.A. Whiting (2004). Convergence of Proportional-Fair sharing algorithms under general conditions. *IEEE Trans. Wireless Commun.* **3**, 1250–1259.

[45] X. Lin, N.B. Shroff (2006). The impact of imperfect scheduling on cross-layer congestion control in wireless networks. *IEEE/ACM Trans. Netw.* **14**, 302–315.

[46] X. Lin, N.B. Shroff, R. Srikant (2008). On the connection-level stability of congestion-controlled communication networks. *IEEE Trans. Inf. Theory* **54**, 2317–2338.

[47] R. Litjens, F. Roijers, J.L. van den Berg, R.J. Boucherie, M.J. Fleuren (2003). Performance analysis of wireless LAN's: an integrated packet/flow level approach. In: *Providing QoS in Heterogeneous Environments, Proc. ITC-18*, Berlin, eds. J. Charzinski, R. Lehnert, P. Tran Gia (North-Holland, Amsterdam), 931–940.

[48] X. Liu, E.K.P. Chong, N.B. Shroff (2003). A framework for opportunistic scheduling in wireless networks. *Comp. Netw.* **41**, 451–474.

[49] J. Liu, A. Proutière, Y. Yi, M. Chiang, H.V. Poor (2007). Flow-level stability of data networks with non-convex and time-varying rate regions. In: *Proc. ACM Sigmetrics 2007*, 239–250.

[50] L. Massoulié (2007). Structural properties of proportional fairness: stability and insensitivity. *Ann. Appl. Prob.* **17**, 809–839.

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

[52] J. Mo, J.C. Walrand (2000). Fair end-to-end window-based congestion control. *IEEE/ACM Trans. Netw.* **8**, 556–567.

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

[54] M.J. Neely, E. Modiano, C.-P. Li (2005). Fairness and optimal stochastic control for heterogeneous networks. In: *Proc. Infocom 2005*.

[55] R. Prakash, V.V. Veeravalli (2002). A time-scale separation technique for the analysis of random access systems with incremental redundancy. In: *Proc. ISIT 2002*.

[56] F. Simatos, D. Tibi (2008). Spatial homogenization in a stochastic network with mobility, http://arxiv.org/abs/0807.1205.

[57] A.L. Stolyar (2005). On the asymptotic optimality of the gradient scheduling algorithm for multi-user throughput allocation. *Oper. Res.* **53**, 12–25.

[58] W. Szpankowski (1988). Stability conditions for multidimensional queueing systems with computer applications. *Oper. Res.* **36**, 944–957.

[59] W. Szpankowski (1994). Stability conditions for some distributed systems: buffered random access systems. *Adv. Appl. Prob.* **26**, 498–515.

[60] L. Tassiulas, A. Ephremides (1992). Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. *IEEE Trans. Aut. Contr.* **37**, 1936–1948.

[61] L. Tassiulas, A. Ephremides (1993). Dynamic server allocation to parallel queues with randomly varying connectivity. *IEEE Trans. Inf. Theory* **30**, 466–478.

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

[63] I.E. Telatar, R.G. Gallager (1995). Combining queueing theory with information theory for multi-access. *IEEE J. Sel. Areas Commun.* **13**, 963–969.

[64] P. Viswanath, D.N.C. Tse, R. Laroia (2002). Opportunistic beamforming using dumb antennas. *IEEE Trans. Inf. Theory* **48**, 1277–1294.