Flow-level performance analysis of some opportunistic scheduling algorithms

Thomas Bonald

A Flow-Level Performance Analysis of some Opportunistic Scheduling Algorithms

T. Bonald
France Telecom R&D
Issy-les-Moulineaux, France
thomas.bonald@francetelecom.com

Abstract— **While fading effects have long been combatted in 2G wireless networks, primarily devoted to voice calls, they are now seen as an opportunity to increase the capacity of 3G networks that incorporate data traffic. The packet delay tolerance of data applications, such as file transfers and Web browsing for instance, allows the system flexibility in scheduling a user's packets. So-called opportunistic scheduling, which ensures transmission occurs when radio conditions are most favorable, is the key component of currently developed systems like cdma2000 IS-856 and UMTS HSDPA. We compare the performance of some scheduling schemes using a flow-level approach where the random evolution of the number of ongoing flows is explicitly taken into account.**

Keywords— **Fast fading, multi-user diversity, opportunistic scheduling, downlink channel, flow-level analysis.**

I. Introduction

Data services are expected to constitute a significant part of traffic in 3G wireless networks. A number of new technologies have recently been standardized to support high data rates and optimize the spectrum utilization of downlink channels. High Data Rate (HDR) systems [5], defined in the 3GPP2 cdma2000 IS-856 standard [3], [15], offer a maximum data rate of 2.4 Mbit/s over a signal bandwidth of 1.25 MHz, while their 3GPP equivalent High Speed Downlink Packet Access (HSDPA) systems [1], [2], [22] offer a maximum data rate of around 10 Mbit/s over a signal bandwidth of 5 MHz. These systems deliver high spectral efficiency by using a TDMA-like strategy with a combination of link adaptation, hybrid ARQ and opportunistic scheduling:

• *Link adaptation* refers to the adaptation of a user's transmission data rate to its radio conditions based on Channel Quality Indicator (CQI) signals sent back by the user to the BS.

• *Hybrid ARQ* allows the transmission of any packet spread over multiple slots to be terminated early, i.e., as soon as the packet is successfully received, so as to adapt the transmission rate to the *actual* radio conditions. This control scheme, based on Chase combining or incremental redundancy, is essential given the errors in channel quality prediction and the necessarily conservative Signal-to-Noise (SNR) thresholds used to ensure a successful transmission.

• *Opportunistic scheduling* seeks to transmit a user's packets in slots when conditions are relatively favorable, based on CQI feedback signals.

These dynamic schemes take advantage of the inherent "elasticity" of data transfers to increase the overall system capacity: instead of wasting radio resources in providing a constant data rate to each user, they dynamically adapt the data rate of each user to optimize the spectrum utilization. The duration of a slot (1.67 ms for HDR systems, $3\times$ 0.67 ms for HSDPA systems) is sufficiently short to benefit from the uncorrelated fast variations of channel quality experienced by active users, the so-called multi-user diversity. Thus fading effects, which have long been combated in 2G wireless networks, are now seen as an opportunity to increase the capacity of 3G wireless networks [19], [26].

The scheduling algorithm is a key component of these time-shared systems. In addition to exploiting multi-user diversity over short time-scales, this algorithm also determines how resources are shared over longer time-scales. An algorithm that always selects the user with the highest CQI is efficient in term of overall throughput but may starve low SNR users, typically located far from the BS. An algorithm that equalizes the data rates of active users, on the other hand, is fair but inefficient as most radio resources are used to sustain the data rate of distant users [7]. A third strategy, which realizes a reasonable trade-off between efficiency and fairness, consists in transmitting to the user with the highest data rate relative to its current mean data rate [14], [26]. The so-called proportional fair scheduler has been studied in [16], [17], [20] and is widely used in currently developed systems. Many other scheduling algorithms have been proposed and analyzed, see e.g. [4], [8], [10], [12], [13], [21], [23], [25].

As a general rule the evaluation of scheduling algorithms is performed with an assumed *static* population of users. We maintain that this may lead to misleading conclusions since the actual set of active users is *dynamic* and varies

as a random process as new data flows are initiated and others complete. In particular, while users are generally assumed to be uniformly distributed in the cell, the location of active users in steady state in fact depends on the scheduling employed. This is due to the inherent elasticity of data transfers: the data rate of any user determines how long that user will stay active. In this paper, we use a flow-level approach where the random evolution of the number of ongoing flows is explicitly taken into account. We compare the performance of two standard algorithms, the maximum SNR scheduler and the proportional fair scheduler, to a new algorithm presented in [8] and referred to as the score-based scheduler.

The flow-level model used in the analysis and the simulations is presented in the next section. In the following two sections, we introduce the notions of cell load and cell capacity and evaluate user performance in terms of flow throughput and blocking rate in a reference scenario where the scheduler is a simple round-robin algorithm. Section V is devoted to the comparison between the maximum SNR scheduler, the proportional fair scheduler and the score-based scheduler. An evaluation of the scheduling gain due to the oppotunistic nature of these algorithms is proposed in Section VI. Section VII concludes the paper.

II. Model

A. Traffic characteristics

Users generate traffic in sessions, each session being composed of a random number of data flows separated by intervals of inactivity. Sessions typically arrive according to a Poisson process, like calls in telephone networks. In any given cell, this results in a flow arrival rate which we denote by λ. Each flow is characterized by its size (in bits). Denoting by σ the mean flow size, we define the traffic intensity in the cell as the product $\lambda \times \sigma$ (in bit/s). This is an exogenous parameter that characterizes the traffic offered to the cell.

Remark 1: The notion of traffic intensity, which is key to the evaluation of cell capacity and user performance, cannot be defined in a static scenario where the number of data flows is assumed to be fixed.

Traffic is not necessarily uniformly distributed in the cell. We will denote by $F(x)$ the "traffic density" function, meaning that flows arrive at rate $\lambda F(x)dx$ in any region of area dx around location x. We have:

$$\int_{\text{cell}} F(x)dx = 1.$$

B. Radio characteristics

Given a fixed number of flows, the data rate of each flow results from the extremely complex interaction between physical phenomena like fading and interference, and transmission control mechanisms like link adaptation, error control and scheduling. This complexity is further increased by the fact that users are mobile. In the following, we characterize the radio environment of a user by its "feasible rate", defined as the data rate it would realize if all radio resources were allocated to it.

Remark 2: The feasible rate of a user may not only be limited by its radio conditions but also by the mobile itself, which may be unable to decode signals with small spreading factors for instance. To avoid the waste of radio resources, a hybrid TDMA/CDMA strategy where several users are multiplexed over the same slot can be used in HSDPA systems [2].

It is worth noting that the feasible rate is time-varying due to fading effects. In particular, adaptive modulation and coding schemes are typically used to adapt the data rate to the instantaneous SNR, based on CQI feedback signals. Fast fading occurring over small time-scales (less than 1 second, say) due to multipath propagation may be exploited by means of opportunistic scheduling (refer to Section V). Slow fading due to shadowing and user mobility, on the other hand, cannot be exploited without compromising the packet delay budget of the users. In the following, we neglect the impact of slow fading. Specifically, we assume that users do not move during data transfers[1]. We denote by $R(x)$ the mean feasible rate of a user located at x. Thus when a simple round-robin scheduler is implemented, the effective data rate of such a user in the presence of n ongoing data flows is $R(x)/n$.

While realistic values of the feasible rate could be determined by means of measurements, it proves more convenient for present purposes to consider the limiting values provided by information theory. It turns out that current practical systems achieve data rates quite close to the theoretical limits. Figure 1 compares the data rates realized by HDR systems [5] to the maximum rate of a Gaussian channel given by Shannon's theorem:

$$W \log_2(1 + \text{SNR}), \tag{1}$$

where W represents the bandwidth, equal to 1.25 MHz for HDR systems, and SNR is the signal-to-noise ratio. We observe that the data rates of the HDR technology are approximately equal to 75% of the Shannon limit.

[1] This has been shown to correspond to a conservative scenario [9]. The mobility of any user in the cell improves its own average performance as well as that of all other users.

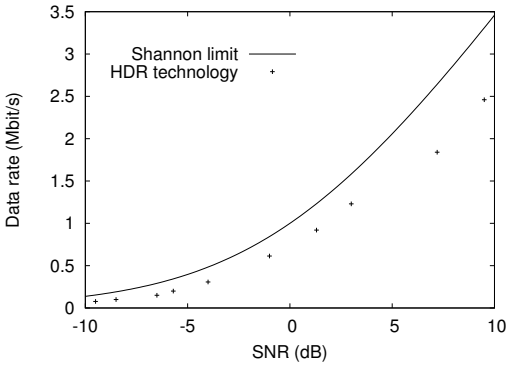

Fig. 1. Efficiency of the HDR technology

In the following, we use the mean feasible rate:

$$R(x) = WE\left[\log_2(1 + \mathrm{SNR}(x))\right], \qquad (2)$$

where $\mathrm{SNR}(x)$ corresponds to the signal-to-interference-plus-noise ratio at location x. This is a random variable due to fading effects, and we assume a perfect link adaptation so that $R(x)$ corresponds to the mean feasible rate averaged over all fading states.

To evaluate the SNR, we consider a homogeneous propagation environment with the following parameters:

- BS transmit power $P = 40\mathrm{dBm}$;
- noise $N = -100\mathrm{dBm}$;
- path loss $\Gamma = 130\mathrm{dBm} + 35\mathrm{dBm} \times \log_{10}(d)$, where d is the distance to the BS (in km).

Note that this corresponds to a path loss exponent $\alpha = 3.5$, with a path loss $\Gamma = 77\mathrm{dBm}$ at distance $d = 30$ m corresponding to free space propagation for a 1 GHz carrier. There is no intracell interference and intercell interference is evaluated in a worst-case scenario where all BSs transmit at full power P.

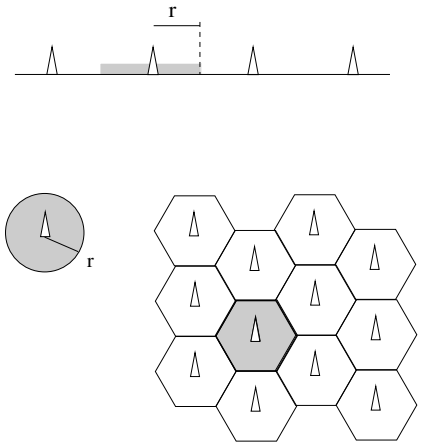

Fig. 2. Linear and hexagonal networks.

For illustrative purposes, we consider linear and hexagonal networks as shown in Figure 2. For linear networks, we refer to the cell radius r as the maximum distance at which a user is served; for hexagonal networks, the cell radius r corresponds to the radius of a disk with the same area as the hexagon. For the evaluation of intercell interference, we consider all BSs whose distance to the reference BS is less than or equal to $10 \times r$, corresponding to 10 interfering BS's for linear networks, 90 interfering BSs for hexagonal networks. We verified that the interference term due to more distant BS's is negligible.

Figure 3 gives the normalized feasible rate $R(x)/W$ with respect to the distance to the BS for linear and hexagonal networks, assuming constant fading. The feasible rate is higher for linear networks due to lower interference.

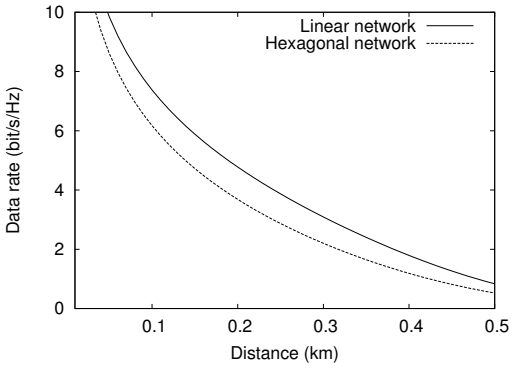

Fig. 3. Feasible rate w.r.t. distance to the BS ($r = 0.5$ km).

In practice, users typically experience either Rayleigh fading, corresponding to a Gaussian distribution of the signal amplitude, or Rician fading when the signal amplitude has a significant line-of-sight component [24]. In this paper, we consider two extreme fading conditions:

- Rayleigh fading, where the amplitude of the signal received from each BS (the BS to which the mobile is attached as well as the interfering BSs) has a Gaussian distribution;
- constant fading, where the amplitude of the signal received from each BS is constant.

A large range of fading conditions can then be generated by varying the proportion of users that experience Rayleigh fading.

Figure 4 shows the impact of Rayleigh fading on the normalized feasible rate $R(x)/W$ for hexagonal networks. We observe a slight increase in the feasible rate due to the fading of interefering signals (the feasible rate (2) is a convex function of interference).

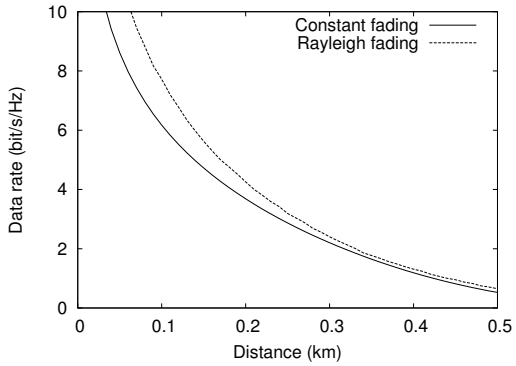

Fig. 4. Feasible rate w.r.t. distance to the BS ($r = 0.5$ km).

III. Cell load, cell capacity

The traffic intensity is an exogenous parameter that characterizes the traffic offered to the cell. The traffic intensity can therefore well exceed the cell capacity in the sense that the flow arrival rate is larger than the flow departure rate. In the absence of admission control, the number of ongoing flows then increases and the data rate of each flow decreases continuously until some users become impatient and interrupt their transfer [11].

We define the cell capacity as the maximum traffic intensity such that the cell is *not* saturated when using a blind scheduling algorithm such as the round-robin scheduler. We first introduce the notion of cell load.

A. Cell load

When using a blind scheduling algorithm, the mean number of slots required to transfer a file to a user located at x is $\sigma/R(x)$, the ratio of the mean file size to the feasible rate. In any region of area dx around location x, data flows arrive at rate $\lambda F(x)dx$. We deduce the infinitesimal load due to any region of area dx around location x:

$$\rho(dx) = \lambda\sigma\frac{F(x)}{R(x)}dx.$$

The cell load is:

$$\rho = \int_{\text{cell}} \lambda\sigma\frac{F(x)}{R(x)}dx. \tag{3}$$

The cell may be viewed as a queueing system of load ρ where the server represents the radio resource (the slots) and the customers the data flows. If $\rho < 1$, the system is stable and the number of customers remains finite; if $\rho > 1$, the system is unstable and the number of customers tends to infinity (in practice, some users abandon their transfer).

B. Cell capacity

We refer to the cell capacity C as the maximum traffic intensity $\lambda\sigma$ such that $\rho < 1$. In view of (3),

$$C = \left(\int_{\text{cell}} \frac{F(x)}{R(x)}dx\right)^{-1}. \tag{4}$$

Note that C corresponds to the weighted *harmonic* mean of the feasible rates $R(x)$, with weights given by the traffic density function $F(x)$, which is less than the corresponding *arithmetic* mean:

$$C = \left(\int_{\text{cell}} \frac{F(x)}{R(x)}dx\right)^{-1} \leq \int_{\text{cell}} F(x)R(x)dx.$$

In particular, the capacity is significantly affected by regions of the cell where the ratio $F(x)/R(x)$ is large, corresponding to a high traffic demand and a low feasible rate. Figure 5 gives the normalized cell capacity C/W for linear and hexagonal networks, with a uniform traffic distribution (constant F) and constant fading. We observe that the cell capacity is maximum for dense networks ($r \to 0$) and tends to zero for sparse networks ($r \to \infty$).

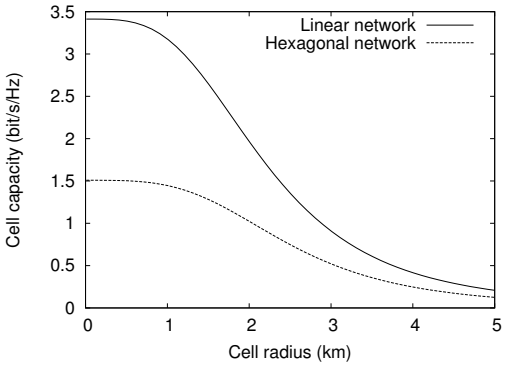

Fig. 5. Cell capacity w.r.t. cell radius (blind scheduler).

Remark 3: We verified that fading variations have an extremely slight impact on the cell capacity. This can be explained by the fact that the cell capacity is mainly determined by the feasible rate at the edge of the cell, which is typically not affected by the fading conditions (refer to Figure 4).

Remark 4: The fact that the capacity of linear networks is much higher than that of hexagonal networks is not only due to lower interference but also to the network topology itself. Consider a user whose location is uniformly distributed in the network. The probability that its distance to the closest BS is less than d is equal to d/r for linear networks, approximately $(d/r)^2$ for hexagonal networks, less than d/r.

The maximum cell capacity obtained for dense networks is interference-limited, i.e., is independent of the noise N and the transmit power P. It only depends on the path loss exponent α. Table I below shows how α impacts the maximum cell capacity. In the rest of the paper, we always take $\alpha = 3.5$.

Cell capacity (bit/s/Hz)	$\alpha = 3$	$\alpha = 3.5$	$\alpha = 4$
Linear network	2.92	3.41	3.84
Hexagonal network	1.22	1.51	1.76

TABLE I
IMPACT OF THE PASS LOSS EXPONENT ON THE MAXIMUM CELL CAPACITY.

IV. USER PERFORMANCE

In this section, we analyse how the cell load impacts user performance in terms of flow throughput and blocking rate in a reference scenario where the scheduler is a simple round-robin algorithm, with and without admission control.

A. No admission control

We define the flow throughput γ as the ratio of the mean flow size σ to the mean flow duration. By Little's law [18], we deduce that the flow throughput is equal to the ratio of the traffic intensity to the mean number of active flows:

$$\gamma = \frac{\lambda\sigma}{E[n]}. \tag{5}$$

For a round-robin scheduler, the number of active flows n evolves like the number of customers in a processor-sharing queue of load ρ. In the absence of admission control, the stationary distribution of n is given by:

$$\pi(n) = \rho^n(1-\rho), \quad \rho < 1.$$

We obtain:

$$E[n] = \frac{\rho}{1-\rho}.$$

Using the fact that

$$\rho = \frac{\lambda\sigma}{C},$$

we deduce from (5) that:

$$\gamma = C(1-\rho).$$

The flow throughput is maximum and equal to the cell capacity C when the cell load is equal to zero, and decreases linearly in the cell load.

Clearly, γ is a mean performance metric, averaged over the cell. One may be interested in a more precise performance metric such as the flow throughput $\gamma(x)$ of users located at x. Let $dn(x)$ be the number of active flows in a region of area dx around x. Such flows arrive at rate $\lambda F(x)dx$. By Little's law, we obtain:

$$\gamma(x) = \frac{\lambda\sigma F(x)dx}{E[dn(x)]}.$$

Using the fact that the expected number of active flows in a region of area dx around x is proportional to the load due to such flows, i.e.,

$$E[dn(x)] = \frac{\rho(dx)}{\rho} \times E[n]$$

with

$$\rho(dx) = \lambda\sigma\frac{F(x)}{R(x)}dx,$$

we deduce:

$$\gamma(x) = R(x)(1-\rho).$$

Thus the flow throughput of users located at x is maximum and equal to the feasible rate $R(x)$ when the cell load is equal to zero, and decreases linearly in the cell load. This is illustrated by Figure 6 using (2) with constant fading and various values of SNR, corresponding to various locations in the cell.

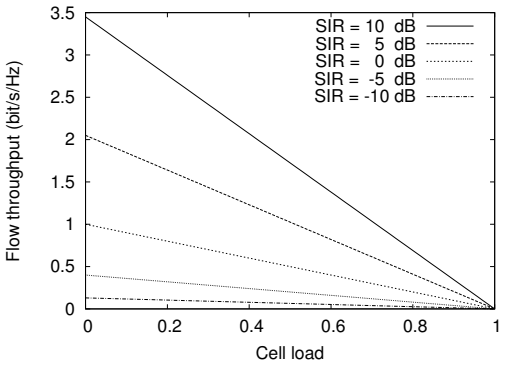

Fig. 6. Flow throughput w.r.t. cell load (round-robin scheduler).

Remark 5: The flow throughput γ corresponds to the weighted *harmonic* mean of the flow throughputs $\gamma(x)$, with weights given by the traffic density function $F(x)$:

$$\gamma = \left(\int_{\text{cell}} \frac{F(x)}{\gamma(x)}dx\right)^{-1}.$$

Remark 6: These results are insensitive to any traffic characteristics (flow size distribution, session structure, etc). The only required assumption is that sessions are independent and arrive as a Poisson process [6].

B. Admission control

Admission control is necessary to offer acceptable data rates in an overload situation where $\rho > 1$. A simple admission policy consists in limiting the number of active flows to a fixed value N. This ensures that the flow throughput $\gamma(x)$ of users located at x is always higher than $R(x)/N$. Quality of service is now perceived not only through the flow throughput but also through the blocking rate. Assuming as in Engset's model that the sessions go on in case of blocking, the stationary distribution of the number of active flows becomes:

$$\pi(n) = \frac{\rho^n}{1 + \rho + \ldots + \rho^N}, \quad n \leq N.$$

We deduce the blocking rate:

$$B = \frac{\rho^N}{1 + \rho + \ldots + \rho^N},$$

and the mean number of active flows:

$$E[n] = \frac{\rho}{1-\rho} \frac{1 + N\rho^{N+1} - (N+1)\rho^N}{1 - \rho^{N+1}}$$

By Little's law, the flow throughput is equal to the ratio of the *actual* traffic intensity to the mean number of active flows, i.e.,

$$\gamma = \frac{\lambda(1-B)\sigma}{E[n]}.$$

Using the fact that

$$\rho = \frac{\lambda\sigma}{C},$$

we obtain:

$$\gamma = C(1-\rho)\frac{1-\rho^N}{1 + N\rho^{N+1} - (N+1)\rho^N}.$$

Similarly, the flow throughput at location x is given by:

$$\gamma(x) = R(x)(1-\rho)\frac{1-\rho^N}{1 + N\rho^{N+1} - (N+1)\rho^N}.$$

Thus the flow throughput at location x decreases from $R(x)$ to $R(x)/N$ as the load ρ goes from 0 to infinity. This is illustrated by Figure 7 for various values of the admission threshold N, using (2) with $\mathrm{SNR} = 0$ dB. The corresponding blocking rate is also shown.

It is worth observing that B decreases exponentially in N at any load $\rho < 1$. If N is sufficiently large, the blocking rate is negligible provided the cell load is not too close to 1. If $N = 100$ for instance, the blocking rate is less than 10^{-3} for a cell load as high as 96%.

If $\rho > 1$, on the other hand, the blocking rate corresponds approximately to the fraction of traffic excess, i.e.,

$$B \sim \frac{\rho - 1}{\rho}.$$

We deduce that if N is sufficiently large (larger than 50, say), the maximum cell load compatible with any reasonable target blocking rate (from 1% to 5%, say) is approximately equal to 1. The maximum traffic intensity is then equal to the cell capacity. This property holds for any admission control scheme, such as an admission decision based on some measure of the current data rate [10], provided new flows cannot be blocked when the number of active flows is too small (less than 50, say).

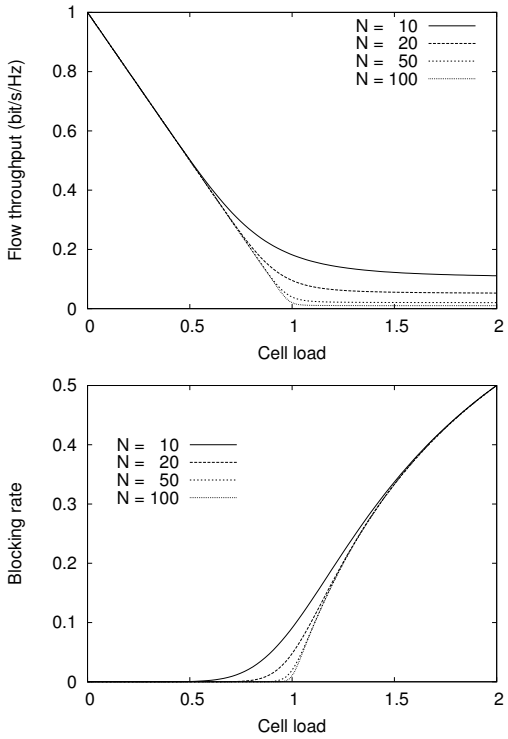

Fig. 7. Flow throughput and blocking rate w.r.t. cell load for various admission thresholds (round-robin scheduler).

V. Opportunistic scheduling

We now study how various opportunistic schedulers impact user performance compared to the previous results derived for a round-robin algorithm. The corresponding flow-level model is a state-dependent multi-class processor-sharing queue which is untractable except under some specific assumptions on the rate statistics [10], [12]. Thus we performed flow-level simulations with Poisson flow arrivals and i.i.d. exponential flow sizes. We first present the considered scheduling algorithms.

A. Maximum SNR scheduler

The simplest algorithm consists in scheduling at each slot the user with the best SNR or, equivalently, the best feasible rate:

$$\text{select} \quad \max\{R_1, \ldots, R_n\}.$$

The main disadvantage of this scheduler is that it tends to always select the same users, those who are close to the BS. In particular, the maximum SNR scheduler does not benefit from the peaks in the feasible rates of the users. This results in reduced efficiency in a dynamic scenario with a varying number of active flows, as will be confirmed by the simulation results.

B. Proportional fair scheduler

This algorithm consists in scheduling at each slot the user with the best feasible rate *relative* to its current throughput:

$$\text{select} \quad \max\left\{\frac{R_1}{T_1}, \ldots, \frac{R_n}{T_n}\right\}.$$

where the throughput $T_i(k)$ of user i at time-slot k is evaluated as follows:

$$T_i(k) = (1 - \frac{1}{t})T_i(k-1) + \frac{1}{t}R_i(k-1) \times 1_{\{i \text{ selected at slot } k-1\}}.$$

The smoothing parameter $1/t$ determines the time constant of the algorithm. A large value of t offers the opportunity of waiting a long time before scheduling a user when its channel quality hits a peak. We then expect the scheduler to better exploit multi-user diversity at the expense of longer packet delays. Thus the time constant should be set accounting for the packet delay tolerance of the applications [26]. For $t = 100$, the typical time-scale of the scheduler is around 100 ms, which leads to delays that are acceptable for most data applications.

C. Score-based scheduler

The algorithm is based on the notion of score [8], with corresponds to the rank of the current feasible rate among the past values observed over a window of fixed size w. The score-based scheduler schedules the user with the best score:

$$\text{select} \quad \min\{S_1, \ldots, S_n\},$$

where the score $S_i(k)$ of user i at time-slot k is defined as:

$$S_i(k) = 1 + \sum_{l=1}^{w-1} 1_{\{R_i(k) < R_i(k-l)\}}.$$

If two users are active for instance and the current feasible rate R_1 of user 1 is the second best rate among its w past rate values while the current feasible rate R_2 of user 2 is the fourth best rate among its w past rate values, user 1 is scheduled, irrespective of the relative values of R_1 and R_2 (refer to Figure 8).

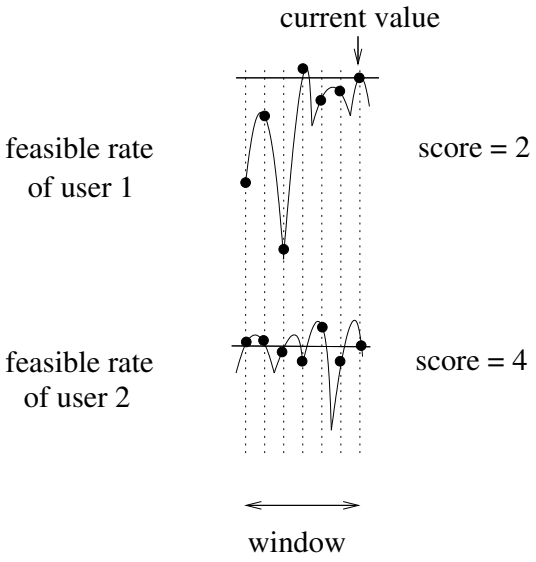

Fig. 8. Principle of the score-based scheduler.

Instead of selecting a user when its feasible rate is high relative to its own *average throughput* (the principle of the proportional fair scheduler), the score-based scheduler selects a user when its transmission rate is high relative to its own *rate statistics*. The corresponding time constant is given by the window size w, which should be set sufficiently large to track the distribution of the rate process while accounting for the packet delay tolerance of applications.

D. Simulation results

We developed a flow-level simulator where at each flow arrival or departure, the data rate of each active flow resulting from the considered algorithm is evaluated. The time-constants of the proportional fair scheduler and the score-based scheduler were set to infinity (the results are approximately insensitive to the time-constants t and w provided these constants are sufficiently large, larger than 100, say). The set of feasible rates is discrete as in practical systems and obtained from (1) with SNR between -20 dB and 20 dB every 1 dB. The maximum number of active flows is fixed to 100. Each simulation run corresponds to 100 000 flow arrivals.

Figure 9 gives the results obtained for a dense hexagonal network with Rayleigh fading (cell radius $r = 0.5$ km). While the flow throughput exhibits approximately the same linear behavior for the three algorithms at low loads, the score-based scheduler outperforms the other two algorithms at high loads. This is confirmed by the results showing the blocking rate at loads close to the critical load. The gain of the score-based scheduler in terms of maximum traffic intensity compatible with any reasonable blocking rate (between 1% and 5%, say) is of 20% over the maximum SNR scheduler, 15% over the proportional fair scheduler.

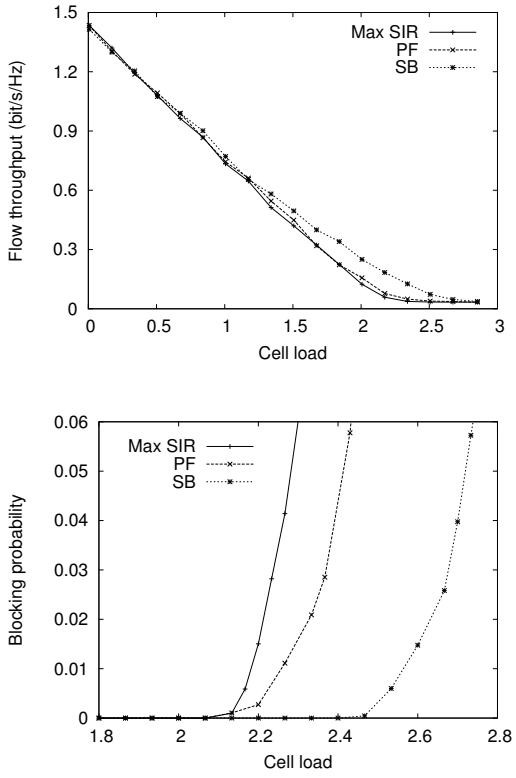

Fig. 9. Flow-level performance for Rayleigh fading.

Figure 10 gives the results obtained when only 50% of users experience Rayleigh fading. The gain of the score-based scheduler in terms of maximum traffic intensity compatible with any reasonable blocking rate is of 15% over the maximum SNR scheduler, 5% over the proportional fair scheduler. Similar results were obtained for linear networks and for various values of the cell radius.

VI. Scheduling gain

In the simulation results of Section V, the critical load where the flow throughput is close to the minimum and the blocking rate becomes non-negligible is much higher than 1. This illustrates the opportunistic nature of the scheduling algorithms compared to a blind round-robin scheduler.

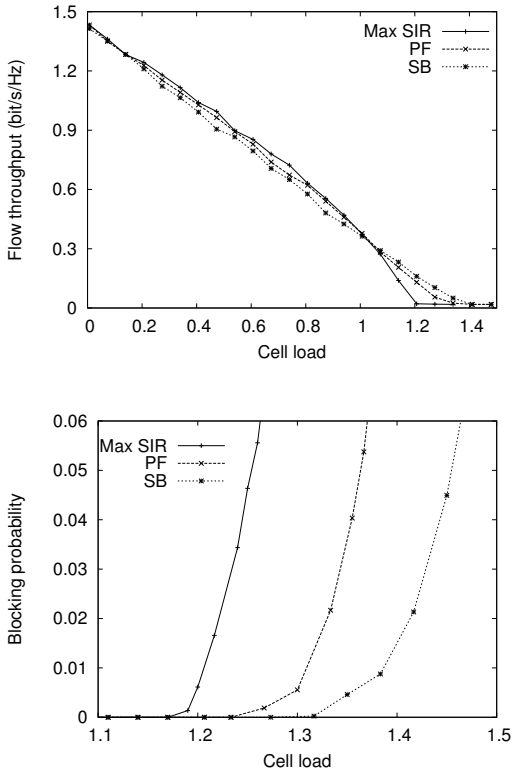

Fig. 10. Flow-level performance for 50% Rayleigh fading.

We now present a method for evaluating this scheduling gain. The idea is to assume that the flow throughput *is* linear in the cell load, as the simulation results suggest, and to evaluate the maximum cell load such that the flow throughput is positive. Thus it is sufficient to evaluate the slope of the throughput vs. load curve at load $\rho = 0$.

A. Homogeneous throughput gain

The slope of the throughput vs. load curve at load $\rho = 0$ is determined by the throughput gain when only two users are active. For sake of clarity, we first assume that this gain H is homogeneous in the sense that it does not depend on the location of these users nor on their radio conditions. The mean number of users is then given by:

$$E[n] = \frac{\rho + 2\rho^2/H}{1+\rho} + o(\rho^2).$$

In view of (5), we obtain:

$$\begin{aligned} \gamma &= C\frac{1+\rho}{1+2\rho/H} + o(\rho) \\ &= C(1 - \frac{2-H}{H}\rho) + o(\rho). \end{aligned}$$

We deduce the scheduling gain:

$$G = \frac{H}{2-H}.$$

If $H = 3/2$ for instance, which is the case for two noise-limited users experiencing Rayleigh fading [8], the scheduling gain is $G = 3$, meaning that the cell capacity is three times larger than that obtained with a round-robin scheduler. This corresponds to the best scheduling gain one might expect, valid for large cells with Rayleigh fading.

B. Heterogeneous throughput gain

In practice, the throughput gain obtained when two users are active does depend on their radio conditions. A key property of the score-based scheduler is that the throughput gain of any user depends on the number of active users and on its own rate statistics only. Thus we evaluate the scheduling gain in this particular case. The result is expected to provide a reasonably accurate estimation of the scheduling gain for the three algorithms.

Let p be the fraction of users with Rayleigh fading. We assume that such a user experiences the same throughput gain H in the presence of another active user, independently of its location in the cell. For Poisson flow arrivals and i.i.d. exponential flow sizes, the stochastic process describing the numbers of active users with Rayleigh fading and with constant fading is a Markov process. It then follows from the balance equations that:

$$E[n] = \frac{\rho + 2\rho^2/\bar{H}}{1+\rho} + o(\rho^2),$$

where $\bar{H}$, which may be viewed as the average throughput gain, is given by:

$$\bar{H} = \left(p^2 + \frac{4p(1-p)}{H+1} + \frac{(1-p)^2}{H} \right)^{-1}.$$

Note that $1 \leq \bar{H} \leq H$ depending on the fraction of Rayleigh fading p. We deduce the scheduling gain as above:

$$G = \frac{\bar{H}}{2 - \bar{H}}.$$

Figure 11 shows the impact of the fraction of Rayleigh fading p on the scheduling gain G for $H = 3/2$. As expected, the scheduling gain increases from 1 to 3 as the fraction of Rayleigh fading increases from 0 to 1.

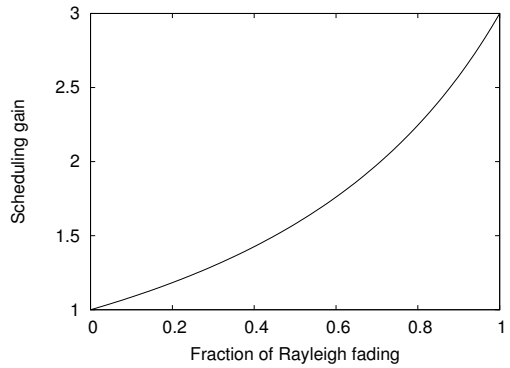

Fig. 11. Impact of fading conditions on the scheduling gain.

VII. Conclusion

We have compared the flow-level performance of two standard algorithms, the maximum SNR scheduler and the proportional fair scheduler, to the score-based scheduler. The simulation results show that the maximum SNR scheduler, the proportional fair scheduler and the score-based scheduler behave similarly at low loads, while the score-based scheduler outperforms the other two algorithms when the cell load is close to critical. The gain in terms of maximum traffic intensity compatible with a given blocking rate is of 15% to 20% over the maximum SNR scheduler, and from 5% to 15% over the proportional fair scheduler, depending on the fading conditions. The efficiency of the score-based scheduler is mainly due to the fact that it fully exploits multi-user diversity by equalizing the slot shares, unlike the other two algorithms that are typically biased against low-SNR users [8].

Another key result is that the flow throughput decreases approximately linearly in the cell load for the three algorithms. Based on this observation, we defined a common notion of scheduling gain, which was shown to vary between 1 and 3 depending on the fading conditions. This means that the cell capacity may be up to three times higher than with a simple round-robin scheduler for pure Rayleigh fading, highlighting the efficiency of opportunistic scheduling and the interest of the HDR evolution of cdma 2000 and the HSDPA evolution of UMTS.

Admission control is a key mechanism to avoid cell saturation and to guarantee a minimum throughput in case of overload. The blocking rate was shown to be negligible while the cell load is not critical, independently of the admission control scheme. The only requirement is that new data flows cannot be blocked when the number of active flows is too small (less than 50, say).

A number of issues should be addressed to complete the comparison between the three considered scheduling algorithms. These include packet-level issues like the packet delay and the interaction with TCP, as well as the ability of the scheduler or some slightly modified version of the scheduler to handle real-time traffic like voice or audio and video streaming.

REFERENCES

[1] 3GPP TS 25.214, "Physical Layer Procedures (FDD)". Release 5, 2003.

[2] 3GPP TS 25.308, "UTRA High Speed Downlink Packet Access (HSDPA); Overall description". Release 5, 2003.

[3] 3GPP2 C.S0024, "cdma2000 High Rate Packet Data Air Interface Specification". 2002.

[4] R. Agrawal, V. Subramanian, "Optimality of certain channel-aware scheduling policies". Proc. 40th Annual Allerton Conf. Commun. Control Comp., 2002.

[5] P. Bender, P. Black, M. Grob, R. Padovani, N. Sindhushayana and A. Viterbi, "CDMA/HDR: A bandwidth-efficient high-speed wireless data service for nomadic users". IEEE Commun. Magazine, 70–77, July 2000.

[6] S. Ben Fredj, T. Bonald, A. Proutière, G. Regnié and J.W. Roberts, "Statistical bandwidth sharing: A study of congestion at flow level". Proc. of ACM SIGCOMM, 2001.

[7] F. Berggren and R. Jäntti, "Asymptotically fair scheduling in fading channels". Proc. of IEEE VTC Fall, 2002.

[8] T. Bonald, " A score-based opportunistic scheduler for fading radio channels". Proc. of European Wireless, 2003.

[9] T. Bonald, S. Borst and A. Proutière, "How mobility impacts the flow-level performance of wireless data systems". Proc. of IEEE INFOCOM, 2004.

[10] T. Bonald and A. Proutière, "Wireless downlink channels: user performance and cell dimensioning". Proc. of ACM MOBICOM, 2003.

[11] T. Bonald and J. Roberts, "Congestion at flow level and the impact of user behaviour", Computer Networks (42) 521–536, 2003.

[12] S. Borst, "User-level performance of channel-aware scheduling algorithms in wireless data networks". Proc. of IEEE INFOCOM, 2003.

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

[14] E.F. Chaponniere, P.J. Black, J.M. Holtzman, D.N.C. Tse, "Transmitter directed code division multiple access system using path diversity to equitably maximize throughput". US Patent 6449490, 2002.

[15] E. Esteves, P.J. Black and M.I. Gurelli, "Link adaptation techniques for high-speed packet data in third generation cellular systems". Proc. of European Wireless Conference, 2002.

[16] J.M. Holtzman, "CDMA forward link waterfilling power control". Proc. of IEEE VTC Spring, 2000.

[17] J.M. Holtzman, "Asymptotic analysis of Proportional Fair algorithm". Proc. of 12th IEEE PIMRC, 2001.

[18] L. Kleinrock, Queueing Systems, Volume I: Theory. Wiley (New York), 1975.

[19] R. Knopp and P. Humblet, Information capacity and power control in single-cell multiuser communications. Proc. of IEEE ICC, 1995.

[20] H.J. Kushner, P.A. Whiting, "Asymptotic properties of Proportional Fair sharing algorithms". Proc. 40th Annual Allerton Conf. Commun. Control Comp., 2002.

[21] X. Liu, E. K.P. Chong, and N.B. Shroff, "A framework for opportunistic scheduling in wireless networks". Computer Networks (41-4) 451–474, 2003.

[22] S. Parkvall, E. Dahlman, P. Frenger, P. Beming, M. Persson, "The high speed packet data evolution of WCDMA". Proc. of the 12th IEEE PIMRC, 2001.

[23] S. Shakkottai, A.L. Stolyar, "Scheduling algorithms for a mixture of real-time and non-real time data in HDR". Proc. of ITC 17, 2001.

[24] B. Sklar, "Rayleigh fading channels in mobile digital communication systems". IEEE Commun. Magazine, 90–100, July 1997.

[25] V. Tsibonis, L. Georgiadis, L. Tassiulas, "Exploiting wireless channel state information for throughput maximization". Proc. of IEEE INFOCOM, 2003.

[26] P. Viswanath, D. Tse and R. Laroia, "Opportunistic beamforming using dumb antennas". IEEE Trans. on Information Theory (48) 1277–1294, 2002.