2E1624 - Performance analysis of communication networks
Project Report

Opportunistic Scheduling in Wireless Networks

Vladimir Vukadinovic and Edouard Drogou
(vvuk@s3.kth.se, drogou@enst.fr)

1. Introduction

In wireless networks, the channel conditions are time-varying due to the fading and shadowing. Different wireless users experience different channel conditions at a given time. This gives raise to the *multi-user diversity* effect: when there are many users which fade independently, at any one time there is a high probability that some of the users will have a strong channel. By allowing only those users to transmit, the shared channel resource is used in the most efficient manner and the total system throughput is maximized. Such scheduling mechanisms are called *opportunistic* because they take advantage of favorable channel conditions in assigning time slots to users. If the service requirements of all the users are flexible, such opportunistic scheduling mechanisms can result in higher spectrum utilization, and increased system throughput. Nevertheless, in practice there are several considerations to take into account before realizing such gains.

To implement the idea of opportunistic scheduling in a real system, two issues need to be addressed: fairness and users' service requirements. In reality, channel statistics of different users are not symmetric and, therefore, a scheme designed only to maximize the overall throughput could be very biased, especially where there are users with widely disparate distances from the base station. For example, allowing only users close to the base station to transmit may result in very high throughput, but sacrifice the transmission of other users. Also, a scheduling strategy should not be concerned only with maximizing long-term average throughputs because, in practice, applications may have different utilities and service constraints. For instance, for real-time applications, the major concern is latency: if the channel variations are too slow, a user may have to wait for a long time before it gets the chance to transmit. When designing a scheduling algorithm, the challenge is to address these issues while at the same time exploiting the multi-user diversity gain inherent in a system. Improving the efficiency of spectrum utilization is important, especially to provide high-rate-data service. However, the potential to exploit higher data throughputs in an opportunistic way, introduces the tradeoff problem between wireless resource efficiency and levels of satisfaction among users.

The cellular system itself also has to satisfy certain requirements in order to extract the multi-user diversity benefits. The base station has to have access to channel quality measurements: in the downlink, each receiver needs to track its own channel signal-to-noise ratio (SNR) and feed back this information to the base station. The base station has to be able to schedule transmissions among the users on a short timescale as well as to adapt users' data rates to the instantaneous channel quality. These features are already present in the designs of many 3.5G high data-rate (HDR) systems. This is the reason why the opportunistic scheduling has received lots of attention recently.

In this project report, we provide a comparative survey of four papers in this area that are listed in the Literature. These papers address the issues that we discussed, either by proposing design

principles for opportunistic scheduling, or by analyzing the performance of existing schemes. In the next Section, we provide a brief survey of objectives and contributions for each paper.

2. Literature overview

First two papers that we survey, [Kushner03] and [Borst05], analyze the performance of the Proportional Fair (PF) scheduler. PF is one of the first opportunistic schedulers that has been proposed in the literature. It has been implemented in CDMA 1xEV-DO system and it has gained considerable attention due to some interesting properties. We give a brief description of the PF algorithm.

Suppose that there are N users in the cell and $R_i(k)$ is *achievable* rate for user i at the transmission interval k, which depends on the user's current channel conditions. Suppose that the scheduler keeps track of the running average rate $\theta_i^{\varepsilon}(k)$ for every user. Then, according to the proportional fair scheduling policy, user $J_k \in \{1,...,N\}$ is chosen for transmission in time slot k if:

$$J_k = \arg\max_{1 \le i \le N} \frac{R_i(k)}{\theta_i^{\varepsilon}(k)}, \quad k = 1,2,... \tag{1}$$

Hence, the user with *relatively* strongest channel is chosen for transmission. The running average rates $\theta_i^{\varepsilon}(k)$ are updated at every time slot as follows:

$$\theta_i^{\varepsilon}(k+1) = \begin{cases} (1-\varepsilon)\theta_i^{\varepsilon}(k) + \varepsilon R_i(k), & i = J \\ (1-\varepsilon)\theta_i^{\varepsilon}(k), & i \neq J_k \end{cases}, \tag{2}$$

where $t_c = 1/\varepsilon$ is the memory of the averaging filter and it is related to the maximum time for which a user can be starved. It can be also observed as the time-scale over which the scheduler aims to provide proportionally fair bandwidth allocation.

It has been shown in **[Kushner03]** that, if the tracking parameter ε in (2) is small and constant, then the path $\theta^{\varepsilon}(\cdot)$ converges weakly to the solution to a deterministic ordinary differential equation (ODE), which is computed from the "mean" dynamics of the throughput process:

$$\dot{\theta}_i = \bar{h}_i(\theta) - \theta_i \ , \tag{3}$$

where $i \le N$ and $\bar{h}_i(\theta) = \mathrm{E}[R_i \cdot J_{\{R_i/\theta_i \ge R_i/\theta_i, j \neq i\}}]$. The path will essentially "follow" the solution to the ODE. The ODE has a unique equilibrium point $\bar{\theta}$ that it is characterized as optimizing a concave utility function:

$$U(\theta) = \sum_i \log \theta_i \ , \tag{4}$$

which shows that PFS is not ad-hoc, but actually corresponds to a concrete maximization problem. The existence of a unique equilibrium is of significance for the performance analysis of the proportional fair algorithm. This is because the equilibrium state will determine the throughput of each user and hence delay. Extensions to multiple antenna and multiple channel systems are also given. Finally, the infinite backlog assumption is dropped and the data is allowed to arrive at random. It has been shown that there is still a mean ODE which characterizes the flow of the algorithm.

In **[Borst05]**, author recognizes that the performance of opportunistic scheduling algorithms has mostly been explored at the packet level for a static user population, often assuming infinite backlogs. Therefore, the paper focuses on the performance at the flow level in a dynamic setting where users come and go as governed by the arrival and completion of random finite-size service demands over time. Two cases have been considered in the paper: symmetric and asymmetric.

In the symmetric case, a scenario with *K* user classes has been considered, where the relative achievable rate fluctuations are statistically identical for all users and the total number of users admitted to the system is *M*. Hence, it has been assumed that the instantaneous rate of user i with time-average rate θ_i is distributed as $R_i \triangleq \theta_i Y_i Z$, where $Y_1, Y_2, ...$ are independent and identically distributed copies and Z represents possible correlation component with unit mean. Class-*k* users submit file transfer requests of size F_k as a Poisson process of rate λ_k. Hence, the offered traffic in class *k* is $\rho_k = \lambda_k \beta_k$, where $\beta_k = \mathrm{E}[F_k / \theta_k]$. The total offered traffic is $\rho = \sum_k \rho_k$. It has been shown that the PF scheduler achieves stability for $\rho < G^*$ or $M < \infty$, where $G^* = \lim_{M \to \infty} G(M)$ and $G(M) = \mathrm{E}[\max Y_j],\ j = 1,...,M$. In this case, the user-level performance 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. Based on the PS model, explicit formulas for the distribution of the number of active users in various classes, mean response times, blocking probabilities, and mean throughput have been provided. It has been also shown that no scheduling strategy achieves stability for $\rho > G^*$.

In the asymmetric case, the relative fluctuations around the respective time-average rates for all users of a given class are statistically identical as before. However, the distributions of the fluctuations are allowed to vary across user classes. Processor-Sharing discipline, which facilitated the analysis in the symmetric case, now becomes largely intractable. Therefore, instead of aiming for full distributional results, the paper focuses on stochastic majorization properties and stability issues of PF scheduler.

Previous two papers analyze the properties of PF scheduler, whose utility function is given by $U_i(\theta) = \log \theta_i$. [Agrawal02] and [Andrews05] consider a more general case of scheduling algorithm design for maximization of an arbitrary, concave increasing utility function.

[Agrawal02] considers a utility maximization problem in the case of multiple channels, i.e., in the case where it is possible to transmit to multiple users at the same time. Assume that there are *d* users in the system. The time-varying channel conditions are captured by a stochastic channel state $\boldsymbol{\eta}_k \in \mathbf{S}$ at time *k*, where $\mathbf{S}$ is the channel state space. Associated with each state $\boldsymbol{\eta} \in \mathbf{S}$ is a rate-region $\mathbf{R}(\boldsymbol{\eta}) \in \Re_+^d$. Thus when the channel is in the state $\boldsymbol{\eta}$, the users may transmit at any vector of rates $\mathbf{v} = (v_1, v_2, ..., v_d) \in \mathbf{R}(\boldsymbol{\eta})$. Also, let $\bar{\mathbf{R}}$ be the set of all achievable steady-state long-term empirical throughput vectors $\mathbf{w}$. The utility maximization problem that [Agrawal02] aims to solve is:

$$U(\mathbf{w}) \triangleq \sup_{\mathbf{w} \in \bar{\mathbf{R}}} \sum_{i=1}^{d} U_i(w_i), \tag{5}$$

where $U_i(\cdot)$ is an increasing, strictly concave and continuously differentiable utility function on $\Re_+$. The following case has been considered: Let $\mathbf{V}_k \in \mathbf{R}(\boldsymbol{\eta}_k)$ be the rate selected at time *k*.

Define $\mathbf{W}_k$ to be the empirical throughput as follows $\mathbf{W}_{k+1} = (1-\varepsilon)\mathbf{W}_k + \mu\mathbf{V}_k$, k>0. Authors show that $\mathbf{W}$ satisfies the following ODE:

$$\dot{\mathbf{W}} = \bar{\mathbf{V}}(\mathbf{W}) - \mathbf{W} \tag{6}$$

A specific choice of scheduling $\mathbf{V}_k(\mathbf{W}_k, \boldsymbol{\eta}_k)$ based on gradient-type algorithm has been analyzed. Three different cases are considered: perfect channel state knowledge, no knowledge of the channel state, and imperfect channel state knowledge. It has been shown that, for instance, perfect channel state knowledge, $\mathbf{V}$ and $\bar{\mathbf{V}}$ have the following forms:

$$\begin{aligned} \mathbf{V}(\mathbf{w}, \boldsymbol{\eta}_k) &= \arg\max_{\mathbf{u}\in\mathbf{R}(\boldsymbol{\eta})} \nabla U(\mathbf{w})^T \mathbf{u} \\ \bar{\mathbf{V}}(\mathbf{w}) &= \arg\max_{\mathbf{u}\in\bar{\mathbf{R}}} \nabla U(\mathbf{w})^T \mathbf{u} \end{aligned} \tag{7}$$

Using properties of the ODE (6) authors show that the algorithm converges to the optimal solution of a related optimization problem.

[Andrews05] introduces an algorithm which seeks to optimize a concave utility function $U(\theta) = \sum_i U_i(\theta_i)$ of the users' throughputs θ_i, subject to certain lower and upper throughput bounds: $\theta_i^{\min} \le \theta_i \le \theta_i^{\max}$. Authors propose an algorithm called the Gradient algorithm with Minimum and Maximum Rate constraints (GMR), which uses a token counter mechanism and has the following form:

$$J(t) = \arg\max_{1\le i\le N} e^{a_i T_i(t)} U_i'(\theta_i(t)) R_i(t), \tag{8}$$

where $\theta_i(t)$ is the current average rate received by user *i*, which is updated as in (2), and $R_i(t)$ is the current achievable rate for user that user. $T_i(t)$ is a *token counter* and $a_i > 0$ is a parameter. The token counter is the key mechanism to enforce the rate constraints. In each time slot it is incremented at rate either $\theta_i^{\min}$ or $\theta_i^{\max}$ and it is decremented whenever the user *i* is served:

$$T_i(t+1) = T_i(t) + R_i^{token} - R_i(t), \tag{9}$$

where $R_i^{token} = \theta_i^{\min}$ if $T_i(t) \ge 0$ and $R_i^{token} = \theta_i^{\max}$ if $T_i(t) < 0$. The dynamics of user throughputs under the proposed GMR algorithm has been studied and it has been shown that GMR is asymptotically optimal in sense that if the throughput vector $\theta(t)$ converges to a fixed vector θ^* as time $t\to\infty$ then θ^* is an optimal solution to the optimization problem described above. Authors also consider two important special cases of the utility functions: $U_i(\theta_i) = \log\theta_i$ and $U_i(\theta_i) = \theta_i$, which correspond to the common proportional fairness and throughput maximization objectives.

3. Overview of analytical methods

In this Section we provide a comparative overview of assumptions and mathematical tools that are used in arriving to the results described above, as well as methods that are used for their verification.

3.1. Assumptions

Even though these four papers are not exactly addressing the same problem, there are some common starting points and assumptions that can be discussed.

Traffic models − In order to evaluate the service received by a user in a system that employs opportunistic scheduling, it is necessary to describe the offered traffic at the flow level. However, scheduling performance has been mostly evaluated at the packet level, assuming that there is an infinite backlog of packets in each queue. Moreover, performance analysis is often performed assuming a static user population. This is also the case in some of the papers that we surveyed, although it is clear that it is not satisfactory to assume that the user population is independent of the scheduling algorithm. For example, a scheduling algorithm that provides high throughput to users with favorable channel conditions will tend to satisfy the service demands of these users sooner. As a result, the algorithm would be left facing a user population with a higher fraction of users with poor channel conditions. In [Agrawal02], a wireless communication system with a constant number of users has been considered, with an implicit assumption on infinite backlog of packets. Also, in [Andrews05] authors assume that, for each flow, there is always data available for service. [Kushner03] goes one step forward: although the infinite backlog assumption is used in deriving the main results in the paper, authors provide an extension of the presented analysis that gives insight in the scheduling performance in the case when users' data arrive at random. No special assumptions on the arrival process have been given. [Borst05] focuses on the performance at the flow level in a dynamic setting with random finite-size service demands. Session arrival processes are assumed to be Poisson. The notion of finite-size service demands allowed the authors in this paper to consider user-perceived performance in terms of response times for file transfers for example, as opposed to delays experienced by individual packets, which is usually done in the case of infinite backlog assumption.

Channel-related assumptions − We address the assumptions on two aspects of the channel: availability of the channel state information and channel access method. Perfect knowledge of the channel state has been often assumed in the literature that studies the performance of opportunistic scheduling. Although 3G systems employ channel estimation and reporting mechanisms, the channel state information available to the base station is not perfect: it is delayed and often outdated. In addition, channel estimation mechanism itself introduces channel estimation errors at the mobile station. [Andrews05], [Kushner03], and [Borst05] assume that the perfect channel knowledge is available at the base station. That practically means that, at every time instance, the opportunistic scheduler knows exactly what the achievable rates for all mobile stations are. The implications of such assumption have not been discussed. [Agrawal02] however, considers three cases: perfect channel knowledge, imperfect knowledge, and no knowledge. Note however, that "no knowledge" still assumes that the information on steady-state achievable data rate for each mobile station is available at the base station. Regarding the channel access method, [Andrews05] and [Borst05] assume time-division multiplex; i.e. only one user is allowed to transmit at a time. The same assumption holds for the most of the analysis presented in [Kushner03], however, an extension to the case when there are multiple channels (or multiple transmit antennas) is also given. [Agrawal02] considers both the case when multiple users are selected for transmission at a time and the TDM case.

Service constraints − No specific service requirements or constraints have been considered in the most the surveyed papers. The only exception is [Andrews05], which considers the simplest possible QoS constraints: minimum and maximum throughput guarantees. Instead of service requirements, main focus of the surveyed papers is on the user's utility. Since [Kushner03] and [Borst05] analyze the performance of the Proportional Fair scheduler, their focus is on the

logarithmic utility function. [Agrawal02] and [Andrews05] consider a more general case of an arbitrary strictly concave increasing utility function.

3.2. Mathematical tools

What all four surveyed papers have in common is that results are presented axiomatically. However, the mathematical tools that are used to arrive to those results are very often quite different.

[Kushner03] recognizes that opportunistic scheduling algorithms are of the stochastic approximation type and it uses the results of stochastic approximation theory to analyze their long-term properties of PF scheduler. The stochastic approximation asymptotic analysis uses continuous time interpolation and weak convergence approach in proving that the steady-state rates of PF scheduler converge to the solution of corresponding ODE. Weak convergence approach is much more flexible than the probability-one method, which is not suitable for the analysis of opportunistic schedulers because of the presence of time-varying parameters. The fact that the solution is unique and globally asymptotically stable is proved by using some results from dynamical systems theory.

[Borst05] also analyzes the PF performance in continuous time. The user dynamics in this paper result from finite-size service demands that arrive randomly over time. It is assumed that the duration of the time slots is short relative to the size and arrival frequency of the service demands. Thus, the scheduling strategy operates on an extremely fast time scale compared to the user dynamics, making it natural to analyze the user-level performance in continuous rather than discrete time, and assume that the users are served simultaneously rather than in a time-slotted fashion. The continuous-time model naturally inherits its service characteristics from the discrete-time model. As we mentioned in the previous Section, [Borst05] considers two cases: symmetric and asymmetric. In the symmetric case, relative fluctuations around the respective time-average rates for all users in the system are statistically identical. Authors show that in this case, the user-level performance of PF may be evaluated by means of a Processor Sharing model. In the asymmetric case, such analysis is not possible. Therefore, in this case, focus is on stochastic majorization properties and stability, rather then on full distributional results.

[Agrawal02], similar to [Kushner03], uses a stochastic approximation approach to show that asymptotically it is possible to analyze the performance of the algorithm by means of an ODE. Using properties of the differential equation authors show that the proposed scheduling algorithms converge to the optimal solution of a related utility maximization problem. Finally, [Andrews05] studies the trajectories of the convergence process using the results from the dynamics of the fluid sample paths (FSP).

3.3. Numerical verification

All four papers provide numerical validation of presented results using simulations. [Kushner03] presents a rather small set of simulation scenarios to illustrate that in the case of PF scheduling, the steady-state throughputs of mobile users converge in time to the solution of the corresponding ODE. Simulation results give almost perfect match. [Borst05] provides an extensive set of simulation scenarios, both for the symmetric and asymmetric case. It assumes a system where users initiate file transfer requests as a Poisson process. Both deterministic and exponentially distributed file sizes have been considered. Results show that the analytical formulas derived in the paper yield a fairly accurate prediction for the mean total number of transfers in progress in the symmetric, as well as in asymmetric case, despite the fact that the rate fluctuations vary across

users. However, the accuracy of the formulas for the mean transfer delays is rather poor in the asymmetric case. The formulas consistently underestimate the delay for the high-SNR users and overestimate the delay for the low-SNR users. [Agrawal02], similar to [Kushner03], provides a comparison of the theoretical results to the solutions of corresponding ODEs. Two rather simplistic simulation scenarios have been presented. In the both scenarios, it has been assumed that there are only two users in the system and that the channel state process takes values in [0, 1] with uniform probabilities. As expected, the trajectories of the different scheduling algorithms proposed in this paper all converge to corresponding ODEs. Finally, [Andrews05] presents extensive simulation results that confirm that the proposed Gradient algorithm with Minimum and Maximum Rate constraints (GMR) is indeed able to satisfy the specified constraints, if feasible. A comparison with the PF scheduler is also included, illustrating that the PF does not provide any type of throughput guarantees. Simulation setup appears to capture many details of the actual transmission environment present in 1xEV-DO system.

4. Conclusion

We surveyed four papers that attempt to put the opportunistic scheduling algorithms on solid mathematical bases. Each of this paper focuses on different aspects of the scheduling problem and represents a piece a 'big picture': [Borst05] addresses the user-level level performance. Users' service constraints are addressed in [Andrews05]. [Agrawal02] and [Kushner03] deal with optimality and convergence properties of opportunistic schemes, respectively. The papers give an insight into a rich set of mathematical tools and methods that can be used in analyzing complex communication problems.

Literature:

[Kushner03] H. J. Kushner and P. A. Whiting, "Convergence of proportional-fair sharing algorithms under general conditions," IEEE Trans. Wireless Communications, 2003.

[Borst05] S. C. Borst, "User-level performance of channel-aware scheduling algorithms in wireless data networks," IEEE Trans. Networking, 2005.

[Agrawal02] R. Agrawal and V. Subramanian, "Optimality of certain channel aware scheduling policies," Allerton Conference on Communication, Control, and Computing 2002.

[Andrews05] M. Andrews, L. Qian, and A. Stolyar, "Optimal utility based multi-user throughput allocation subject to throughput constraints," INFOCOM 2005.