Centrum voor Wiskunde en Informatica

REPORTRAPPORT

PNA

Probability, Networks and Algorithms

Probability, Networks and Algorithms

Inter-cell scheduling in wireless data networks

T. Bonald, S.C. Borst, A. Proutière

REPORT PNA-R0408 DECEMBER 2004

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

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

Probability, Networks and Algorithms (PNA)

Software Engineering (SEN)

Modelling, Analysis and Simulation (MAS)

Information Systems (INS)

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

ISSN 1386-3711

Inter-cell scheduling in wireless data networks

ABSTRACT
Over the past few years, the design and performance of channel-aware scheduling strategies have attracted huge interest. In the present paper we examine a somewhat different notion of scheduling, namely coordination of transmissions among base stations, which has received little attention so far. The inter-cell coordination comprises two key elements: (i) interference avoidance; and (ii) load balancing. The interference avoidance involves coordinating the activity phases of interfering base stations so as to increase transmission rates. The load balancing aims at diverting traffic from heavily-loaded cells to lightly-loaded cells. We consider a dynamic scenario where users come and go over time as governed by the arrival and completion of random data transfers, and evaluate the potential capacity gains from inter-cell coordination in terms of the maximum amount of traffic that can be supported for a given spatial traffic pattern. We also show that simple adaptive strategies achieve the maximum capacity without the need for any explicit knowledge of the traffic characteristics. Numerical experiments demonstrate that inter-cell scheduling may provide significant capacity gains, the relative contribution from interference avoidance vs. load balancing depending on the configuration and the degree of load imbalance in the network.

2000 Mathematics Subject Classification: 60K25, 68M20
Keywords and Phrases: Wireless data networks; downlink; inter-cell scheduling; load balancing; stability; network capacity

Inter-Cell Scheduling in Wireless Data Networks

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

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

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

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

***Abstract*— Over the past few years, the design and performance of channel-aware scheduling strategies have attracted huge interest. In the present paper we examine a somewhat different notion of scheduling, namely coordination of transmissions among base stations, which has received little attention so far. The inter-cell coordination comprises two key elements: (i) interference avoidance; and (ii) load balancing. The interference avoidance involves coordinating the activity phases of interfering base stations so as to increase transmission rates. The load balancing aims at diverting traffic from heavily-loaded cells to lightly-loaded cells. We consider a dynamic scenario where users come and go over time as governed by the arrival and completion of random data transfers, and evaluate the potential capacity gains from inter-cell coordination in terms of the maximum amount of traffic that can be supported for a given spatial traffic pattern. We also show that simple adaptive strategies achieve the maximum capacity without the need for any explicit knowledge of the traffic characteristics. Numerical experiments demonstrate that inter-cell scheduling may provide significant capacity gains, the relative contribution from interference avoidance vs. load balancing depending on the configuration and the degree of load imbalance in the network.**

***Index Terms*— Wireless data networks, downlink, inter-cell scheduling, load balancing, stability, network capacity.**

I. Introduction

Wireless networks are evolving to support a wide variety of high-speed data applications, in addition to conventional voice services and current low-bandwidth data services such as short messaging. Data applications tend to have drastically different traffic characteristics and QoS requirements than voice connections, calling for fundamentally different resource allocation mechanisms. In particular, wireless circuit-switched voice networks rely on power control algorithms for adjusting the transmit power so as to compensate for the varying channel quality and maintain a fixed transmission rate. Various data applications on the other hand, such as file transfers and Web browsing sessions, do not have a stringent rate requirement and are less sensitive to packet-level delays. Such *elastic* applications are well-suited for rate control algorithms which adapt the transmission rate over time to track the fluctuations in the channel quality while transmitting at constant (maximum) power.

Adaptive rate control mechanisms offer the possibility to improve the throughput performance by scheduling the data transmissions and exploiting the relative delay tolerance of data users. A particularly attractive approach, in fading environments, is to schedule the transmissions to the various users when their channel conditions are (relatively) favorable, as in the Proportional Fair algorithm for the CDMA 1xEV-DO system [9], [21]. While fading is considered to have a predominantly adverse impact for voice connections, it thus provides the opportunity to achieve throughput gains for elastic data transfers. The design and performance of such channel-aware or *opportunistic* scheduling algorithms have been extensively studied over the past few years. Most of the studies have focused on the performance at the packet level for a static user population [2], [3], [16], [24], [27], [30], [31], although recently the flow-level performance in a dynamic setting has been analyzed as well [10], [15].

In the present paper we focus on a different notion of scheduling, namely coordination of transmissions among base stations (BS's), which has received relatively little attention so far. The inter-cell scheduling that we consider comprises two key elements: (i) *interference avoidance*; and (ii) *load balancing*. The rationale for *interference avoidance* stems from the simple fact that the feasible transmission rates are impacted by the amount of interference from surrounding BS's, so that significantly higher rates may be achieved when neighboring BS's are switched off. When the increase in the feasible rates is sufficiently large, it may outweigh the sacrifice of transmission resources at the BS's that are turned off, yielding a net benefit. Note that the increase in the rates is especially substantial for users on the cell boundary which are affected the most by interference from neighboring BS's. Since the edge users are the most demanding ones in terms of transmission resources, they have a major impact on the overall performance.

The above observations are confirmed by the results in [8], [20], [26] which show that such inter-cell scheduling strategies achieve substantial throughput gains in a static scenario with a fixed ensemble of users. They are also indirectly supported by the findings in [11] which examines the flow-level performance of networks of BS's in a dynamic setting. The latter paper does actually not consider inter-cell scheduling (the BS's are assumed to be on as long as there are any users to be served). However, it shows that accounting for the time-varying activity patterns of surrounding BS's makes a crucial difference, suggesting potential scope for significant gains from coordinating the transmission activities of interfering BS's.

As mentioned above, inter-cell coordination is different in nature from opportunistic scheduling. However, the active control of interference is somewhat related in the sense that coordinating the activity phases of interfering BS's may be interpreted as a form of network-wide opportunistic scheduling by generating favorable channel conditions in a coordinated fashion. Antenna-based incarnations of the latter concept (opportunistic beam-forming to artificially induce channel variations) have been explored in [31].

The motivation for *load balancing* arises from the natural principle that the overall performance may be improved by diverting traffic from heavily-loaded BS's to lightly-loaded BS's. This may be achieved by allocating users to BS's not solely on the basis of signal strength measurements, but taking load considerations into account as well, either average values or instantaneous conditions, see for instance [12], [19]. As the latter already implies, shifting traffic in wireless networks is hindered by the fact that signal strengths and feasible transmission rates tend to vary widely among candidate BS's, so that moving traffic from the strongest received BS to weaker ones comes at the expense of raising the total load. In particular, it is generally not optimal to perfectly balance the loads. The above penalty is limited however for edge users which do not have a strong affinity with a nearest BS. Again, such users are also the most onerous ones in terms of resource requirements, creating a promising opportunity for performance improvements.

In the present paper we examine the potential capacity gains from inter-cell coordination in a situation where users come and go over time as governed by the arrival and completion of random data transfers. Congestion manifests itself in this context by the number of active users competing for access to the transmission resources. In particular, the network may be unstable in the sense that the number of users may grow indefinitely. Thus we introduce as in [14] the notion of *network capacity* as the maximum amount of traffic compatible with stability for some given spatial traffic pattern. In order to elucidate the coordination aspects, we consider the simplest possible scenario with no radio channel variations due to shadow fading or multi-path propagation.

In the subsequent analysis we focus on a scenario where each user is uniquely attached to a serving BS and each BS, when active, transmits to a single user. Such networks are representative of the CDMA 1xEV-DO system for instance, and will be simply referred to as 'TDMA networks'. Numerical experiments demonstrate that inter-cell scheduling may provide significant capacity gains in this practically interesting case, the relative contribution from interference avoidance vs. load balancing depending on the configuration and the degree of load imbalance in the network.

As a final comment, it is worth observing that yet a further degree of inter-cell coordination may be accomplished by transmitting to users from several BS's simultaneously and essentially operating the BS's as an antenna array. While the structure of the mathematical models that we consider is general enough to cover such approaches as well, the derivation of the feasible transmission rates for such schemes poses an (information-theoretic) problem by itself, and moves beyond the scope of the present paper.

The remainder of the paper is organized as follows. In Section II we present a detailed model description, and describe how networks of coordinated BS's may be modeled in the general context of queueing systems with interacting service resources. In Section III we examine the stability region and introduce the notion of network capacity. We also show that simple adaptive scheduling strategies achieve the maximum capacity. These results are applied to 'TDMA networks' in Section IV. Section V is devoted to the numerical experiments and Section VI concludes the paper.

II. Model description

We consider the downlink of a network of BS's $\mathcal{N} = \{1, \ldots, N\}$ whose transmission resources (power, bandwidth, codes, etc) are shared by a dynamic population of data flows. Flows arrive at random and leave the network once the corresponding data transfer has been completed. Each flow is characterized by its size (in bits) and its data rate that depends on the user's location, which is assumed to be fixed throughout the entire flow duration, and the way the transmission resources are allocated among active flows. Without loss of generality, we consider a set of classes $\mathcal{I} = \{1, \ldots, I\}$ such that flows of any given class have the same size statistics and rate characteristics, as described below. We denote by x_i the number of class-i flows and by $x = (x_1, \ldots, x_I)$ the network state.

A. Traffic characteristics

The traffic model is allowed to be very general. We simply assume that class-i flows arrive as a stationary and ergodic marked point process of intensity λ_i, with marks corresponding to the flow sizes. Let σ_i be the mean size of class-i flows (in bits). The traffic intensity of class i is then defined by $\rho_i = \lambda_i \times \sigma_i$ (in bits/s). We denote by $\rho = (\rho_1, \ldots, \rho_I)$ the traffic intensity vector, by $\varrho = \sum_{i=1}^{I} \rho_i$ the total traffic intensity, and by $p_i = \rho_i / \varrho$ the proportion of the total traffic intensity generated by class i.

B. Resource allocation

The service capabilities of the network are described by a set of available transmission 'profiles' $\mathcal{J} = \{1, \ldots, J\}$. Each of the profiles corresponds to a particular allocation of the transmission resources among the various classes. At any point in time, one of the available profiles can be selected for operating the network.

When profile j is selected, the class-i flows share a data rate $R_{i,j}$, which is equal to zero when class i is not served in profile j. Evidently, the set of available profiles and the corresponding service rates strongly depend on the degree of flexibility in deploying the transmission resources. For now, however, we do not make any specific assumptions on the service rates, nor are we concerned how exactly the service rate of a class is shared among the active flows.

When profile j is used a fraction of the time α_j, the resulting service rate of class i is $r_i = \sum_{j \in \mathcal{J}} \alpha_j R_{i,j}$. We assume that the switching between transmission profiles can be executed with sufficiently fine time granularity compared to the flow dynamics so that the flows experience a constant service rate r_i.

The time allocation $\alpha = (\alpha_1, \ldots, \alpha_J)$ is termed the *scheduling strategy*. In the following, we denote by $\mathcal{T}$ the set of non-negative vectors α such that $\sum_{j \in \mathcal{J}} \alpha_j = 1$. When the time fractions are independent of the network state, the scheduling strategy is referred to as *static*. The strategy is called *adaptive* when the time fractions *do* depend on the network state. The set of achievable rate vectors $r = (r_1, \ldots, r_I)$ is referred to as the rate region:

$$\mathcal{R} = \left\{ r : \ \exists \alpha \in \mathcal{T}, \ \forall i \in \mathcal{I}, \ r_i \leq \sum_{j \in \mathcal{J}} \alpha_j R_{i,j} \right\}.$$

Remark 2.1: The above-described model falls in the general framework of queueing systems with interacting service resources. Such models were introduced in the early nineties for evaluating the performance of parallel systems and wireless networks [7], [29], and have more recently been analyzed in [4], [5], [6], [28].

III. Stability region and network capacity

We are interested in determining the stability region of the network, i.e., the set of traffic intensity vectors ρ such that there exists a scheduling strategy for which the network state x has a finite stationary distribution. In addition, we are interested in finding *stable* scheduling strategies, i.e., scheduling strategies that stabilize the network whenever possible. We first characterize the stability region and discuss the stability of some scheduling strategies. We then introduce the key notion of *network capacity*, defined for given traffic proportions of the classes $(p_1, \ldots, p_I)$ as the maximum traffic intensity ϱ compatible with stability.

A. Stability region

The stability region has been characterized in [4], [5]. It coincides with the interior of the rate region $\mathcal{R}$.

Proposition 3.1: There exists a scheduling strategy for which the network is stable if and only if $\rho \in \breve{\mathcal{R}}$, with:

$$\breve{\mathcal{R}} = \left\{ r : \ \exists \alpha \in \mathcal{T}, \ \forall i \in \mathcal{I}, \ r_i < \sum_{j \in \mathcal{J}} \alpha_j R_{i,j} \right\}.$$

The necessary condition follows trivially from the observation that the carried traffic vector must belong to the interior of the rate region. The sufficient condition readily follows from the fact that if $\rho \in \breve{\mathcal{R}}$, then there exists a time allocation α such that:

$$\forall i \in \mathcal{I}, \quad \rho_i < \sum_{j \in \mathcal{J}} \alpha_j R_{i,j}.$$

Under that static strategy, the network behaves as a set of I independent stable queues.

B. Stability of some scheduling strategies

1) Static scheduling: We just observed that for any vector in the stability region there exists a static scheduling strategy α for which the network is stable. For any given static scheduling strategy α, on the other hand, the network is stable if and only if $\rho_i < r_i$ for all $i \in \mathcal{I}$, with $r_i = \sum_{j \in \mathcal{J}} \alpha_j R_{i,j}$. The corresponding stability region is a strict subset of $\breve{\mathcal{R}}$, except in the trivial case where the set $\mathcal{J}$ reduces to a single transmission profile. We conclude that static scheduling strategies are not stable, in contrast to some adaptive strategies presented below.

When the traffic proportions among classes $(p_1, \ldots, p_I)$ are fixed, however, there exists a unique static scheduling strategy α that stabilizes the network whenever possible. It is obtained by solving the linear program:

$$\begin{aligned} &\text{minimize} && \sum_{j \in \mathcal{J}} \tau_j && (1) \\ &\text{subject to} && \tau_j \geq 0, \ \sum_{j \in \mathcal{J}} \tau_j R_{i,j} \geq p_i. \end{aligned}$$

The solution of this linear program τ^* corresponds to the minimum amount of time needed to transmit p_i bits of class i, for all $i \in \mathcal{I}$. Thus the maximum total traffic intensity is $1/\tau^*$, and the fraction of time that profile j is used is $\alpha_j = \tau_j^*/\tau^*$.

2) Traffic-based scheduling: A natural adaptive scheduling strategy consists in selecting the transmission profile based on traffic variables like the workload or the number of active flows. Armony & Bambos [4], [5] studied a particular class of traffic-based scheduling strategies that they refer to as ‘MaxProjection’ schedulers where a profile j is selected that corresponds to the maximum rate vector $(R_{1,j}, \ldots, R_{I,j})$ in the direction of the workload vector. They proved that such a scheduling strategy is stable. Since the workload is generally not known in practice, we consider a slightly modified strategy which may be interpreted as a continuous-time analogue of the ‘MaxWeight’ strategy in [22] where the scheduling decision is based on the number of flows. Specifically, in state x a profile j is selected that maximizes:

$$\sum_{i \in \mathcal{I}} V(x_i) R_{i,j}, \qquad (2)$$

where $V(\cdot)$ denotes any increasing, strictly concave function on $\mathbb{R}_+$. The next proposition is proved in the appendix for a fairly general class of functions described in Mo & Walrand [25].

Proposition 3.2: The traffic-based scheduling strategy (2) is stable.

3) Utility-based scheduling: A further adaptive scheduling strategy is based on the notion of utility which has been widely adopted in the context of wireline networks [23]. Let $U(\cdot)$ be an increasing, strictly concave function on $\mathbb{R}_+$ representing per-flow utility. For any given state x, the scheduling strategy is such that the rate vector r is a solution of the optimization problem:

$$\begin{aligned} &\text{maximize} && \sum_{i \in \mathcal{I}} x_i U(r_i/x_i) && (3) \\ &\text{subject to} && r \in \mathcal{R}. \end{aligned}$$

The next proposition is proved in the appendix for the class of utility functions introduced in [25].

Proposition 3.3: The utility-based scheduling strategy (3) is stable.

Note that the maximization problem (3) is not easy to solve in general, unlike (2) which can be solved through a simple scan of the transmission profiles. Under time-slotted operation, the ‘gradient’ scheduling algorithm introduced by Stolyar [28] provides a means for solving (3) that reduces to a scan of the transmission profiles in every time slot. Specifically, in time slot t, the gradient algorithm selects a transmission profile $j(t)$ and a set of k_i class-i flows $\mathcal{K}_i(t) \subseteq \{1, \ldots, x_i\}$ that maximize

$$\sum_{i \in \mathcal{I},\ k \in \mathcal{K}_i} U'(T_{i,k}(t)) \times R_{i,j}/k_i,$$

where $T_{i,k}(t)$ denotes the exponentially smoothed throughput received by the k-th class-i flow, given for some fixed coefficient w, $0 < w \leq 1$, by:

$$T_{i,k}(t+1) = (1-w)T_{i,k}(t) + w 1_{k \in \mathcal{K}_i(t)} R_{i,j}/k_i.$$

For any fixed network state, this algorithm converges to the solution of the optimization problem (3) as $t \to \infty$ and $w \downarrow 0$.

C. Network capacity

Given some fixed traffic proportions of the various classes $(p_1, \ldots, p_I)$, it is natural to define the network capacity as the maximum admissible total traffic intensity ϱ, i.e., the maximum value of C such that the network is stable whenever $\varrho < C$. By virtue of Proposition 3.1, the network capacity can be expressed as the optimal value of the following optimization problem:

$$\begin{aligned} &\text{maximize} && C \\ &\text{subject to} && (p_1, \ldots, p_I) C \in \mathcal{R}. \end{aligned}$$

In view of §III-B.1, the solution C^* of this optimization problem is equal to $1/\tau^*$, where τ^* is the solution of the linear program (1). Thus evaluating the network capacity is equivalent to finding the optimal *static* scheduling strategy.

IV. Inter-cell scheduling in TDMA networks

In the analysis of the stability region and the network capacity we allowed the set of available profiles to be completely general, providing great flexibility in deploying the transmission resources. In the remainder of the paper we will focus on interference avoidance in a specific scenario where (i) each class is associated with a unique serving BS and (ii) each BS, when active, transmits to a single user at full power. We will fix the traffic proportions of the various classes, and then examine the gain in terms of network capacity resulting from interference avoidance.

It is worth observing that, although it is sufficient to focus on the optimal static scheduling strategy as determined by (1) to evaluate the network capacity, such a scheduling strategy is not stable and therefore vulnerable from a practical perspective. In view of Propositions 3.2 and 3.3, the same capacity gains are achievable through more robust, adaptive strategies such as traffic-based or utility-based scheduling algorithms.

We now specify the model in greater detail, then proceed to consider the illustrative case of a two-cell network and derive the network capacity for a class of symmetric networks.

A. Transmission profiles for TDMA networks

As mentioned above, we assume each class to be associated with a unique serving BS. Let $\mathcal{I}_n = \{1, \ldots, I_n\}$ be the set of flow classes served by BS n. The i-th class served by BS n is referred to as class ni, $i \in \mathcal{I}_n$. Denote by p_{ni} the proportion of the total traffic intensity generated by class ni.

We further assumed that each BS, when active, transmits to a single user at full power. Thus, any transmission profile j is determined by a set of active BS's $\mathcal{A}$ and a set of classes $\{i_n \in I_n; n \in \mathcal{A}\}$ served by the active BS's. Denote by $C_{ni,\mathcal{A}}$ the 'feasible' rate of any class-ni flow when the set of active BS's is $\mathcal{A} \subseteq \mathcal{N}$, i.e., the effective data rate of such a flow when served by BS i and the set of active BS's is $\mathcal{A}$. By convention, $C_{ni,\mathcal{A}} = 0$ if $n \notin \mathcal{A}$. The service rate of class ni when profile j is used is then $R_{ni,j} = C_{ni,\mathcal{A}}$ if $n \in \mathcal{A}$ and $i = i_n$, $R_{ni,j} = 0$ otherwise.

B. A two-cell network

We first consider the illustrative example of a network with just two BS's $\mathcal{N} = \{1, 2\}$. For compactness, denote by $C_{ni,\text{on}} \equiv C_{ni,\{1,2\}}$ and $C_{ni,\text{off}} \equiv C_{ni,\{n\}}$ the feasible rate of class-ni flows when the other BS is on and off, respectively.

1) No interference avoidance: In order to determine the capacity gain from interference avoidance, we first evaluate the network capacity in the absence of such a mechanism, i.e., each of the BS's with at least one active flow transmits at full power, independently of the network state. Such networks have been considered in [11]. In case of two BS's and a single class per BS, the model corresponds to a coupled-processors system [17]. Let

$$\gamma_n = \sum_{i \in \mathcal{I}_n} \frac{\rho_{ni}}{C_{ni,\text{on}}},$$

and define $\bar{C}_{1i} = \gamma_2 C_{1i,\text{on}} + (1 - \gamma_2) C_{1i,\text{off}}$ if $\gamma_2 < 1$, $\bar{C}_{2i} = \gamma_1 C_{2i,\text{on}} + (1 - \gamma_1) C_{2i,\text{off}}$ if $\gamma_1 < 1$. Assuming that the capacity is fairly shared among active flows, the stability condition then reads:

$$\gamma_1 < 1 \quad \text{and} \quad \sum_{i \in \mathcal{I}_2} \frac{\rho_{2i}}{\bar{C}_{2i}} < 1,$$

or

$$\gamma_2 < 1 \quad \text{and} \quad \sum_{i \in \mathcal{I}_1} \frac{\rho_{1i}}{\bar{C}_{1i}} < 1.$$

We derive the network capacity as in §III-C. In case of homogeneous loads, $\gamma \equiv \gamma_1 = \gamma_2$, the stability condition reduces to $\gamma < 1$. The network capacity then equals $2c$, where c denotes the per-cell capacity:

$$c = \sum_{i \in \mathcal{I}_n} \frac{p_{ni}}{C_{ni,\text{on}}}.$$

2) Interference avoidance: The network capacity resulting from interference avoidance is given by $1/\tau^*$, with τ^* denoting the solution of the linear program (1). In case of two BS's, this linear program reads as follows:

$$\begin{aligned} \text{minimize} \quad & \tau = \tau_{\text{on}} + \tau_{1,\text{off}} + \tau_{2,\text{off}} \qquad (4) \\ \text{subject to} \quad & C_{ni,\text{on}} \tau_{ni,\text{on}} + C_{ni,\text{off}} \tau_{ni,\text{off}} \geq p_{ni} \\ & \quad i \in \mathcal{I}_n,\ n = 1, 2, \\ & \sum_{i=1}^{I_1} \tau_{1i,\text{on}} = \sum_{i=1}^{I_2} \tau_{2i,\text{on}} = \tau_{\text{on}} \\ & \sum_{i=1}^{I_1} \tau_{1i,\text{off}} = \tau_{1,\text{off}},\ \sum_{i=1}^{I_2} \tau_{2i,\text{off}} = \tau_{2,\text{off}} \\ & \tau_{ni,\text{on}}, \tau_{ni,\text{off}} \geq 0 \quad i \in \mathcal{I}_n, n = 1, 2, \end{aligned}$$

with $\tau_{ni,\text{on}}$ and $\tau_{ni,\text{off}}$ representing the amount of time that class ni is served at BS n while the other BS is on and off, respectively.

To solve the above linear program, observe that classes with a large ratio $C_{ni,\text{off}}/C_{ni,\text{on}}$ enjoy significantly higher rates when the other BS is switched off. Thus, it is advantageous to serve such classes when the other BS is inactive, and serve the classes with a smaller ratio when both BS's are active. This is confirmed by the next proposition. Without loss of generality, we assume that the flow classes are indexed in such a way that:

$$\frac{C_{n1,\text{off}}}{C_{n1,\text{on}}} \le \frac{C_{n2,\text{off}}}{C_{n2,\text{on}}} \le \ldots \le \frac{C_{nI_n,\text{off}}}{C_{nI_n,\text{on}}}, \quad n = 1, 2. \qquad (5)$$

Proposition 4.1: There exist a solution of (4) and $i_n \in \{1, \ldots, I_n + 1\}$, $n = 1, 2$, such that the following two properties hold: (i) $\tau_{ni,\text{off}} = 0$ for all $i < i_n$, and (ii) $\tau_{ni,\text{on}} = 0$ for all $i > i_n$.

Proof. Let $i_n = \max\{i' : \tau_{ni,\text{off}} = 0 \text{ for all } i < i'\}$, then property (i) follows by definition.

To prove property (ii), assume that $\tau_{ni,\text{on}} > 0$ for some $i > i_n$. By definition, $C_{ni,\text{on}}$ bits of class i are transmitted when BS n serves class i during one second when the other BS is active as well. If BS n serves class i_n instead, then $C_{ni_n,\text{on}}$ bits are transmitted. Since $\tau_{ni_n,\text{off}} > 0$, we may assume that these bits are transmitted while the other BS is active in the original scheduling scheme. Therefore, an amount of time $C_{ni_n,\text{on}}/C_{ni_n,\text{off}}$ becomes available for transmitting to class i while the other BS is inactive. The number of transmitted bits of class i is then $C_{ni_n,\text{on}}/C_{ni_n,\text{off}} \times C_{ni,\text{off}}$, which is larger than or equal to $C_{ni,\text{on}}$ by assumption (5). Thus, either $\tau_{ni,\text{on}}$ or $\tau_{ni_n,\text{off}}$ may be reduced to zero without increasing the total amount of transmission time, from which the desired statement follows. □

Let i_1, i_2 be indices satisfying the properties (i) and (ii) of Proposition 4.1. It then follows that

$$\tau_{\text{on}} = \sum_{i \le i_1} \tau_{1i,\text{on}} = \sum_{i \le i_2} \tau_{2i,\text{on}}$$

and

$$\tau_{1,\text{off}} = \sum_{i \ge i_1} \tau_{1i,\text{off}}, \quad \tau_{2,\text{off}} = \sum_{i \ge i_2} \tau_{2i,\text{off}},$$

with

$$\forall i < i_n,\ \tau_{ni,\text{on}} = \frac{p_{ni}}{C_{ni,\text{on}}}, \quad \forall i > i_n,\ \tau_{ni,\text{off}} = \frac{p_{ni}}{C_{ni,\text{off}}}.$$

It remains to determine i_n and $\tau_{ni_n,\text{on}}, \tau_{ni_n,\text{off}}$, $n = 1, 2$, which follows from the above equalities and the next proposition.

Proposition 4.2: There exists a solution of (4) such that

$$\tau_{1i_1,\text{on}}\tau_{2i_2,\text{on}} = 0 \quad \text{and} \quad \frac{C_{1i_1,\text{on}}}{C_{1i_1,\text{off}}} + \frac{C_{2i_2,\text{on}}}{C_{2i_2,\text{off}}} \le 1,$$

or $i_1 = I_1 + 1$ or $i_2 = I_2 + 1$. In addition, one of the following three properties holds:

(i) $\tau_{1i_1,\text{on}} = \tau_{2i_2,\text{on}} = 0$ and either $i_1 = i_2 = 1$ or $i_1 > 1,\ i_2 > 1,\ \dfrac{C_{1i_1-1,\text{on}}}{C_{1i_1-1,\text{off}}} + \dfrac{C_{2i_2-1,\text{on}}}{C_{2i_2-1,\text{off}}} \ge 1;$

(ii) $\tau_{1i_1,\text{on}} > 0,\ \tau_{2i_2,\text{on}} = 0$ and $i_2 > 1$,
$$\frac{C_{1i_1,\text{on}}}{C_{1i_1,\text{off}}} + \frac{C_{2i_2-1,\text{on}}}{C_{2i_2-1,\text{off}}} \ge 1;$$

(iii) $\tau_{1i_1,\text{on}} = 0,\ \tau_{2i_2,\text{on}} > 0$ and $i_1 > 1$,
$$\frac{C_{1i_1-1,\text{on}}}{C_{1i_1-1,\text{off}}} + \frac{C_{2i_2,\text{on}}}{C_{2i_2,\text{off}}} \ge 1.$$

Proof. We prove the first assertion only. The proof of the second assertion follows in a similar fashion.

Let i_1, i_2 be maximum indices satisfying the properties (i) and (ii) of Proposition 4.1. Assume that $i_1 \le I_1$ and $i_2 \le I_2$. We then have $\tau_{ni_n,\text{off}} > 0$ for $n = 1, 2$. Hence, it is not disadvantageous to serve classes i_n sequentially rather than simultaneously. Thus we can choose either $\tau_{1i_1,\text{on}} = 0$ or $\tau_{2i_2,\text{on}} = 0$. In addition, $C_{ni_n,\text{on}}$ bits of class i_n would be transmitted if BS's n served classes i_n simultaneously during one second. When each BS n individually serves class i_n instead when the other BS is inactive, transmitting these bits requires an amount of time $C_{1i_1,\text{on}}/C_{1i_1,\text{off}} + C_{2i_2,\text{on}}/C_{2i_2,\text{off}}$. We deduce that the latter quantity does not exceed 1. □

It is worth observing that all classes are served in a single profile, except for class i_1 in case (ii) and class i_2 in case (iii). In symmetric conditions $\mathcal{I}_1 = \mathcal{I}_2$, $p_{1i} = p_{2i}$, $C_{1i,\text{on}} = C_{2i,\text{on}}$ and $C_{1i,\text{off}} = C_{2i,\text{off}}$, all classes are served in a single profile.

C. Symmetric networks

We have characterized the optimal scheduling strategy for two-cell networks using the class ordering (5). For networks with more than two cells, classes cannot be ordered in such a way and the optimal scheduling strategy becomes very difficult to characterize. We now define a class of symmetric networks for which the network capacity can be easily evaluated.

We consider a network with a finite number of BS's $\mathcal{N} = \{1, \ldots, N\}$. The approach has a straightforward generalization for an infinite number of BS's, as shown in the examples of Section V. We implicitly consider modulo-N indices so that BS 0 coincides with BS N.

Definition 1 (Symmetric network): We say that a network is symmetric if the BS's serve equivalent sets of classes $\mathcal{I}_0 \equiv \mathcal{I}_1 = \ldots = \mathcal{I}_N$ and these classes have the same rate and traffic characteristics in the sense that for all $i \in \mathcal{I}_0,\ \mathcal{A} \subseteq \mathcal{N},\ n \in \mathcal{A}$:

$$C_{ni,\mathcal{A}} = C_{0i,\mathcal{A}-n}, \quad \rho_{ni} = \rho_{0i}.$$

Thus all cells are equivalent for a symmetric network. In the following, we consider a reference BS, say BS 0, and simply write

$$C_{i,\mathcal{A}} \equiv C_{0i,\mathcal{A}}, \quad \rho_i \equiv \rho_{0i}, \quad p_i \equiv p_{0i}.$$

1) No interference avoidance: In the absence of interference avoidance, the stability condition can be easily derived due to the fact that all cells are equally loaded [11]. It simply reads:

$$\sum_{i \in \mathcal{I}_0} \frac{\rho_i}{C_{i,\mathcal{N}}} < 1.$$

We obtain the network capacity as Nc, where c denotes the per-cell capacity without interference avoidance:

$$c = \left(\sum_{i \in \mathcal{I}_0} \frac{p_i}{C_{i,\mathcal{N}}} \right)^{-1}.$$

2) Interference avoidance: In order to evaluate the capacity gain from interference avoidance, we make the additional assumption that the admissible transmission profiles are symmetric in the sense of the definitions below. This assumption ensures that the optimal static scheduling strategy has a simple structure. We verified in the examples of Section V that asymmetric transmission profiles do not further improve the network capacity.

Definition 2 (Symmetric set of BS's): We say that a set $\mathcal{A} \subseteq \mathcal{N}$ containing the reference BS is symmetric if there exists a permutation σ of $\mathcal{I}_0$ such that for all $i \in \mathcal{I}_0$ and $k \in \mathbb{Z}$:

$$C_{i,\mathcal{A}} = C_{\sigma^k(i),\mathcal{A}-n_k}, \quad \rho_i = \rho_{\sigma^k(i)}, \tag{6}$$

where $\{n_k, \ k \in \mathbb{Z}\}$ denotes the successive elements of $\mathcal{A}$, with $n_0 = 0$. We refer to $\{\sigma^k(i), \ k \in \mathbb{Z}\}$ as the set of classes 'associated' with i.

Definition 3 (A base of symmetric sets of BS's): We say that the sets $\mathcal{A}_l \subseteq \mathcal{N}$, $l \in \mathcal{L}$, constitute a base of symmetric sets of BS's if there exists a permutation σ of $\mathcal{I}_0$ such that (6) is satisfied for all sets $\mathcal{A}_l$, $l \in \mathcal{L}$.

Consider the transmission profiles generated by the sets of active BS's $\{\mathcal{A}_l - n, \ l \in \mathcal{L}, \ n \in \mathcal{N}\}$, where the sets $\mathcal{A}_l \subseteq \mathcal{N}$, $l \in \mathcal{L}$, constitute a base of symmetric sets of BS's. The network capacity can be evaluated using the following algorithm:

For each class $i \in \mathcal{I}_0$:

1) Evaluate $C^*_{i,l} = \max_{n \in \mathcal{A}_l} C_{i,\mathcal{A}_l - n}$ for each $l \in \mathcal{L}$;
2) Evaluate $C^*_i = \max_{l \in \mathcal{L}} d_l \times C^*_{i,l}$, where d_l denotes the *density* of the set $\mathcal{A}_l$, i.e., the number of elements of $\mathcal{A}_l$ divided by N.

Proposition 4.3: The network capacity is equal to Nc^*, where c^* denotes the per-cell capacity:

$$c^* = \left(\sum_{i \in \mathcal{I}_0} \frac{p_i}{C^*_i} \right)^{-1}.$$

Proof. For any $i \in \mathcal{I}_0$ and $l \in \mathcal{L}$, each active BS in the set $\mathcal{A}_l$ can serve either class i or a class associated with i at the maximum rate $C^*_{i,l}$. By symmetry, the optimal scheduler uses the sets of active BS's $\mathcal{A}_l - n$, $n \in \mathcal{N}$, the same fraction of time. Since each BS is active a fraction of time d_l, we deduce that the set of classes associated with i can be fairly served at the maximum rate $d_l \times C^*_{i,l}$. Since all classes associated with i have the same traffic intensity, an optimal scheduling strategy consists in serving these classes in a single transmission profile l, that that maximizes $d_l \times C^*_{i,l}$. The corresponding rate is C^*_i and the minimum time required to transmit p_i bits of class i is given by p_i / C^*_i. We deduce that the minimum time required to transmit p_i bits of class i for all $i \in \mathcal{I}_0$ is given by:

$$\tau^* = \sum_{i \in \mathcal{I}_0} \frac{p_i}{C^*_i}.$$

□

V. Numerical experiments

We now apply the previous results to evaluate the potential capacity gain due to interference avoidance in TDMA networks with canonical topologies like linear and hexagonal networks. We first specify the propagation model. Unless stated otherwise, we assume a uniform traffic distribution across the entire network.

A. Radio environment

We consider a continuous setting where the feasible rate of a user located at x is $C_{x,\mathcal{A}}$ when the set of active BS's is $\mathcal{A}$. We assume that the feasible rate depends on the signal-to-noise ratio (SNR) through the Shannon formula:

$$C_{x,\mathcal{A}} = W \log_2 (1 + \mathrm{SNR}_{x,\mathcal{A}}),$$

where W represents the bandwidth and $\mathrm{SNR}_{x,\mathcal{A}}$ is the SNR of a user located at x when the set of active BS's is $\mathcal{A}$. The Shannon formula provides a reasonable approximation of most real systems, up to a multiplicative constant. Let y_n be the location of BS n. For any x in the cell of the reference BS, say BS 0, we get for isotropic radio propagation:

$$\mathrm{SNR}_{x,\mathcal{A}} = \frac{P\Gamma(|x - y_0|)}{N_0 + P \sum_{n \in \mathcal{A}, n \neq 0} \Gamma(|x - y_n|)},$$

where P denotes the common transmit power of the BS's, N_0 is the background noise level and Γ is the path loss. For the numerical results, we take values representative of 3G wireless networks: $P = 40$ dBm,

$N_0 = -100$ dBm, $\Gamma(r) = -130 - 35\log_{10}(r)$ dBm with r in kilometers. Note that this corresponds to a path loss exponent of 3.5.

Remark 5.1: The path loss $\Gamma(r)$ becomes larger than 1 for very small values of r, say $r < 30$ cm. We verified that the precise value of the path loss at such small distances does not affect the network capacity.

The above model is valid for non-directional antennas, which will be the default assumption. For directional antennas, the path loss Γ is multiplied by a function $H(\theta)$ representing the antenna gain in direction θ, the angle to the center of the beam. We then take the following standard function [1]:

$$H(\theta) = -\min\left\{12\left(\frac{3\theta}{2\pi}\right)^2, 20\right\} \text{ dBm}.$$

B. Coordinating two BS's

We first consider the case of two BS's, separated by a distance $2R$ and serving users located on the segment between these BS's. Each user is served by the closest BS. Classes are indexed by the distance r to the BS, $r \le R$. We denote by $C_{r,\text{on}}$ and $C_{r,\text{off}}$ the feasible rate of a user at distance r of the BS when the other BS is on and off, respectively. In view of Proposition 4.2, the optimal static scheduler serves users at distance r when the opposite BS is on if and only if $r < r^*$, where r^* is defined by $C_{r^*,\text{on}} = C_{r^*,\text{off}}/2$ (see Figure 1).

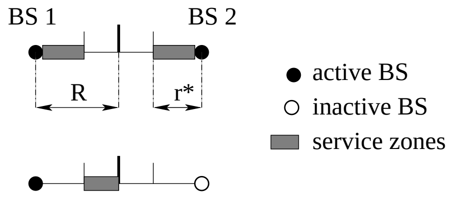


Fig. 1. Optimal static scheduling strategy for a two-cell symmetric network, $R = 0.5$ km.

The cell capacity is then given by:

$$c^* = \left(\int_0^{r^*} \frac{dr}{C_{r,\text{on}}} + 2\int_{r^*}^{R} \frac{dr}{C_{r,\text{off}}}\right)^{-1}.$$

Figure 2 compares the cell capacity obtained with and without interference avoidance.

Remark 5.2: The gain due to interference avoidance can be extremely large for small cells. This is due to the fact that the ratio $C_{r,\text{off}}/C_{r,\text{on}}$ is very high for all $r \le R$. More generally, it is most often advantageous to switch on BS's one at a time in a very dense network with a finite number of BS's.

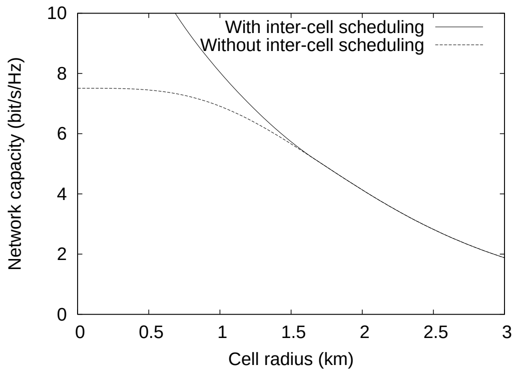


Fig. 2. Capacity of a two-cell symmetric network.

Now assume traffic is not uniformly distributed, but proportional to the distance to BS 1, so the total traffic load in cell 2 is three times that in cell 1. Besides the cell capacity obtained with and without interference avoidance, we are interested in the impact of load balancing where the cell radii R_1, R_2, with $R_1 + R_2 = 2R$, are set to equalize the cell loads:

$$\int_0^{R_1} \frac{rdr}{C_{r,\text{on}}} = \int_{R_1}^{R_1+R_2} \frac{rdr}{C_{r,\text{on}}}.$$

The corresponding cell capacity with and without interference avoidance can be derived from the results of §IV-B. We observe in Figure 3 that load balancing increases the capacity of large cells where interference has a limited effect. On the other hand, load balancing is inefficient for small cells where interference is strong, and can even have a negative impact in this case. This is due to the fact that load balancing tends to maximize the use of the transmission resources, that is to use all BS's at the same time, which in turn maximizes interference.

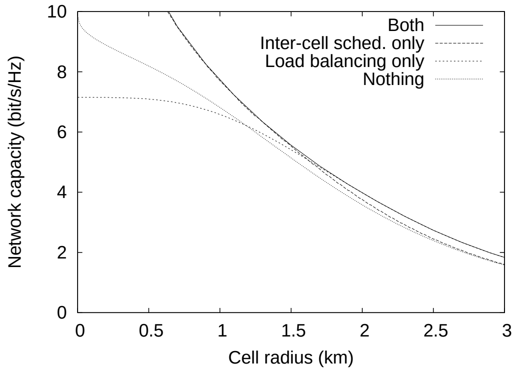


Fig. 3. Capacity of a two-cell asymmetric network.

This example highlights the fact that load balancing and interference avoidance are somewhat contradictory schemes. While load balancing tends to make all BS's active to maximize the use of their transmission resources, interference avoidance aims at forcing some

BS's to be idle to limit the impact of interference on edge users. Thus, while load balancing is most efficient in sparse, noise-limited networks, interference avoidance is most efficient in dense, interference-limited networks.

C. Coordinating three BS's

Consider now a network of three BS's as represented in Figure 4. This is a symmetric network with classes indexed by (r, θ), the distance r to the BS and the angle θ to the center of the beam. A base of symmetric sets of BS's is given by $\mathcal{A}_1 = \{1, 2, 3\}$, $\mathcal{A}_2 = \{1, 2\}$, $\mathcal{A}_3 = \{1\}$, with respective densities $d_1 = 1$, $d_2 = 2/3$, $d_3 = 1/3$, and permutation $\sigma(r, \theta) = (r, -\theta)$.

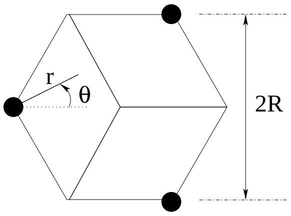


Fig. 4. A three-cell symmetric network.

Figure 5 gives the optimal static scheduling strategy, i.e., the transmission profiles and the corresponding service zones for $R = 0.5$ km and $R = 0.2$ km (up to symmetry relations, i.e., the set of all transmission profiles is obtained by symmetry). Figure 6 compares the cell capacity obtained with and without interference avoidance.

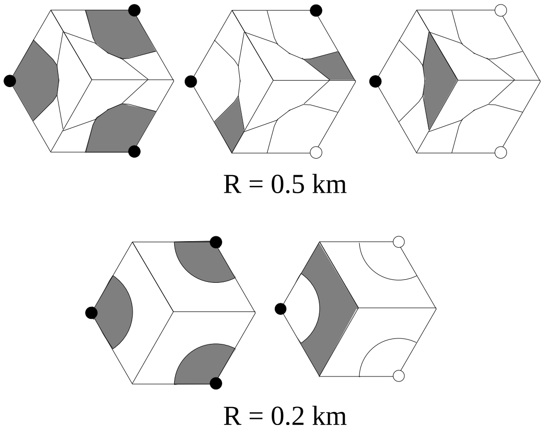


Fig. 5. Optimal static scheduling strategy for a three-cell network.

D. Coordinating three facing sectors in tri-sectorized hexagonal networks

We now consider an infinite tri-sectorized hexagonal network and we quantify the gain resulting from the coordination of three facing sectors. The only difference with the three-cell network lies in the interference term. For the sake of simplicity, we assume that all the BS's except the three considered BS's are always active. We consider a directional antenna with a beam centered at angle $\theta = 0$, see Figure 4.

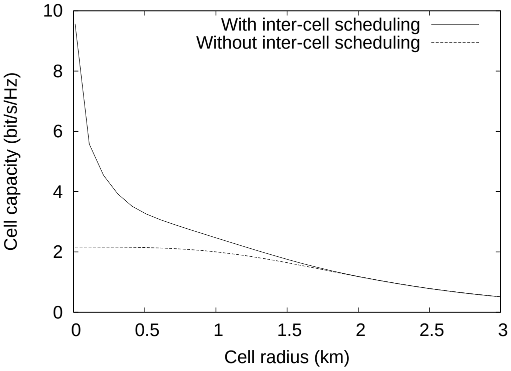


Fig. 6. Capacity of a three-cell network.

Figure 7 gives the optimal static scheduling strategy, (again, up to symmetry relations). We verified that the optimal scheduling strategy is roughly insensitive to the cell radius for sufficiently small cells, $R < 0.5$ km say.

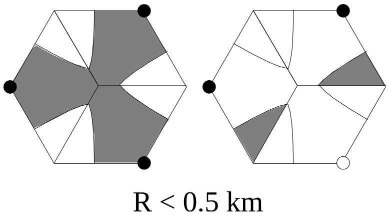


Fig. 7. Optimal static scheduling strategy for a tri-sectorized hexagonal network.

Figure 8 compares the cell capacity obtained with and without interference avoidance. The capacity gain may be as high as 33% for dense networks, a high value given the limited degree of coordination between the BS's.

E. Linear networks

Consider now an infinite linear network as depicted in Figure 9. Two successive BS's are separated by a distance $2R$.

This is a symmetric network with classes indexed by $r \in (-R, R)$. The sets $\mathcal{A}_{1j} = \{jk,\ k \in \mathbb{Z}\}$, $j \geq 1$, and $\mathcal{A}_{2j} = \{jk, jk+1, k \in \mathbb{Z}\}$, $j \geq 3$, form a base of symmetric sets of active BS's with respective densities $d_{1j} = 1/j$, $d_{2j} = 2/j$ and permutation $\sigma(r) = -r$.

It turns out that the optimal static scheduling strategy uses two transmission profiles only: that where all BS's

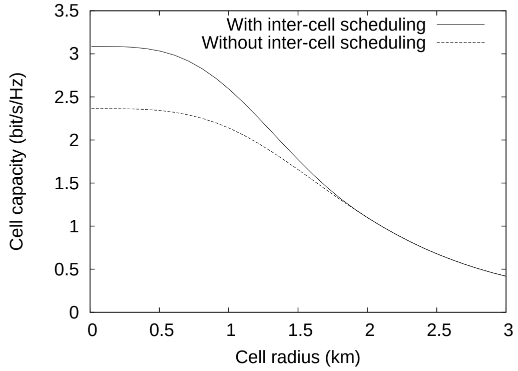


Fig. 8. Capacity of a tri-sectorized hexagonal network.

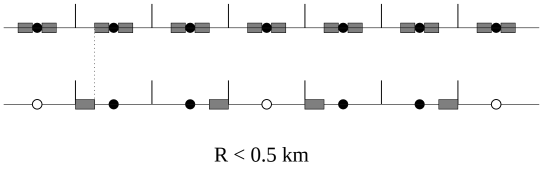


Fig. 9. Optimal static scheduling strategy for a linear network.

are active, corresponding to the set $\mathcal{A}_{11}$, and that where the closest interfering BS is turned off, corresponding to the set $\mathcal{A}_{23}$. A class r is served when all BS's are on if and only if $r < r^*$, with r^* such that $C^*_{r,\mathcal{A}_{11}} = 2/3 \times C^*_{r,\mathcal{A}_{23}}$. We observed that the ratio r^*/R is almost constant and approximately equal to 0.54 for sufficiently small cells, $R < 0.5$ km say. Figure 9 shows the corresponding scheduling strategy, Figure 10 gives the cell capacity with and without interference avoidance. The capacity gain may be as high as 50% for dense networks.

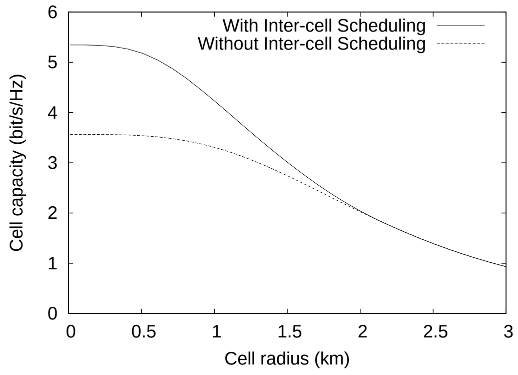
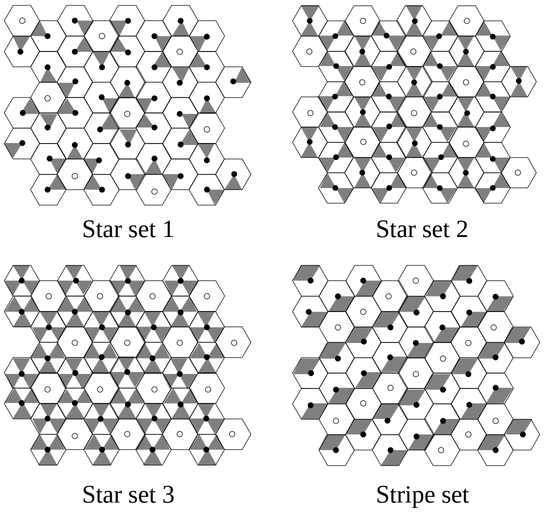


Fig. 10. Capacity of a linear network.

F. Hexagonal networks

Finally, we consider an infinite hexagonal network. This is a symmetric network with classes indexed by (r, θ) as in §V-C. All 'connex' components of any symmetric set of BS's have the same number of elements, possibly infinite. When finite, one can verify that this number is equal to 1, 2, 3 or 6, in which case the connex components are singletons, two neighboring BS's, triangles and hexagons, respectively. If infinite, the symmetric set is either the whole network, a 'stripe' set or one of the three 'star' sets depicted in Figure 11, where the stripe set of highest density is also represented. The permutation defining the base of symmetric sets is $\sigma(r, \theta) = (r, \theta + \pi/3)$.


Star set 1

Star set 2

Star set 3

Stripe set

Fig. 11. Star and stripe sets of respective densities 6/7, 3/4, 2/3, 2/3 and the corresponding potential service zones.

Again, it turns out that the optimal static scheduling strategy uses two transmission profiles only: that where all BS's are active and that where the closest interfering BS is turned off, corresponding to the star set 1. The optimal scheduling strategy is again roughly the same for sufficiently small cells, and represented in Figure 12.

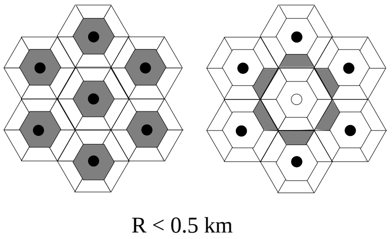


Fig. 12. Optimal static scheduling strategy for a hexagonal network.

Figure 13 gives the cell capacity with and without interference avoidance. The capacity gain may be as high as 25% for dense networks.

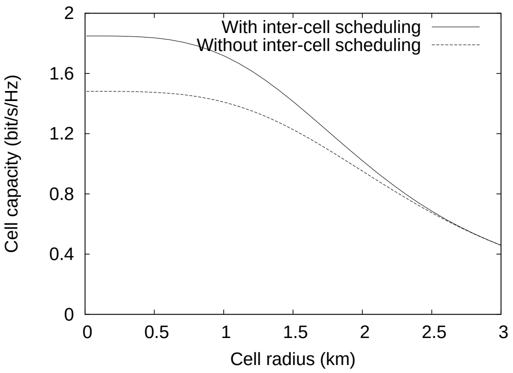


Fig. 13. Capacity of a hexagonal network.

VI. Conclusion

We have evaluated the potential capacity gains from inter-cell scheduling in a dynamic setting, and specifically focused on interference avoidance in ‘TDMA networks’. Numerical experiments indicate that interference avoidance yields the largest gains in dense, interference-limited networks, ranging from 25 to 50%, depending on the network topology. In sparse, noise-limited networks, the gains from interference avoidance alone are far smaller, but load balancing *does* help increase capacity in case of heterogeneous loads. Thus, the relative efficacy of interference avoidance vs. load balancing critically depends on the network configuration.

Another key observation is that, in all considered examples, the optimal capacity is attained by the use of two transmission profiles only: that where all BS are on, and that where only the first interfering BS is switched off. This suggests that a limited degree of coordination is sufficient in practice to achieve the expected capacity gains. The design of the corresponding distributed scheduling schemes opens research perspectives of considerable practical interest.

Appendix

In this appendix, we prove Propositions 3.2 and 3.3 by means of fluid limit techniques [18]. For convenience, we assume Poisson flow arrivals and i.i.d. exponentially distributed flow sizes so that the network state $x(t)$ at time t constitutes a Markov process. The proof can be easily extended to more general traffic characteristics [18].

Proof of Proposition 3.2

As in [4], we consider the fluid limits obtained when the initial number of flows grows to infinity:

$$X(t) = \lim_{\omega\to\infty} \frac{x(\omega t)}{\omega} \quad \text{with} \sum_{i=1}^{I} x_i(0) = \omega.$$

Denote by $t_j(t)$ the cumulative amount of time that profile j is used in the interval $(0,t)$ and define:

$$T_j(t) = \lim_{\omega\to\infty} \frac{t_j(\omega t)}{\omega}.$$

The evolution of the class-i ‘fluid’ volume $X_i(t)$ is determined by the following differential equation. For all t such that $X_i(t) > 0$,

$$\frac{dX_i}{dt} = \lambda_i - \frac{1}{\sigma_i} \times \sum_{j\in A(X)} R_{i,j} \frac{dT_j}{dt},$$

where $A(x) = \arg\max_{j\in\mathcal{J}} \sum_{i\in\mathcal{I}} V(x_i) R_{i,j}$. Note that the set $A(x)$ is not necessarily restricted to a single profile j. By continuity, we have:

$$\sum_{j\in A(X)} \frac{dT_j}{dt} = 1,$$

and for all $\beta \in \mathcal{T}$,

$$\sum_{i\in\mathcal{I}} \sum_{j\in\mathcal{J}} V(X_i) R_{i,j} \beta_j \le \sum_{i\in\mathcal{I}} \sum_{j\in A(X)} V(X_i) R_{i,j} \frac{dT_j}{dt}.$$

Now assume $V(\cdot)$ belongs to the class of utility functions introduced by Mo & Walrand [25]:

$$V(x) = \frac{x^{1-\beta}}{1-\beta}, \quad \beta > 0, \beta \neq 1,$$

and consider the Lyapunov function:

$$f(x) = \sum_{i\in\mathcal{I}} V(x_i) \sigma_i x_i.$$

If $\rho \in \breve{\mathcal{R}}$, then there exists $\theta < 1$ and $\beta \in \mathcal{T}$ such that:

$$\rho_i = \theta \sum_{j\in\mathcal{J}} \beta_j R_{i,j}.$$

We deduce:

$$\begin{aligned}
\frac{df(X)}{dt} &= \gamma \sum_{i\in\mathcal{I}} V(X_i) \frac{dX_i}{dt} \\
&= \gamma \sum_{i\in\mathcal{I}} V(X_i) \left(\rho_i - \sum_{j\in A(X)} R_{i,j} \frac{dT_j}{dt} \right) \\
&\le \gamma(\theta - 1) \sum_{i\in\mathcal{I}} \sum_{j\in\mathcal{J}} V(X_i) \beta_j R_{i,j} \\
&\le \gamma'(\theta - 1) \sum_{i\in\mathcal{I}} V(X_i) \\
&\le \gamma''(\theta - 1) f(X)^{\gamma'''},
\end{aligned}$$

for some constants $\gamma, \gamma', \gamma'', \gamma'''$. We conclude that there exists a finite t_0 such that $f(X(t)) = 0$ for all $t > t_0$. The fluid system is stable, which implies the ergodicity of the Markov process $x(t)$. □

Proof of Proposition 3.3

Consider utility functions of the form:

$$U(x) = \frac{x^{1-\beta}}{1-\beta}, \quad \beta > 0, \beta \neq 1,$$

as introduced by Mo & Walrand [25]. The proof follows as in [13] by introducing the Lyapunov function:

$$f(x) = \sum_{i \in \mathcal{I}} \sigma_i \rho_i^{-\beta} \frac{x_i^{\beta+1}}{\beta + 1}.$$

Note that the result can be extended to more general utility functions as in [32]. □

References

[1] 3GPP (2002). Spatial Channel Model Text Description - SCM 077. http://www.3gpp.org.

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

[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] M. Armony (1999). *Queueing Networks with Interacting Service Resources.* Ph.D. thesis, Stanford University.

[5] M. Armony, N. Bambos (1999). Queueing networks with interacting service resources. In: *Proc. 37th Annual Allerton Conf. Commun., Control, Comp.*, 42–51.

[6] N. Bambos, G. Michailidis (2004). Queueing and scheduling in random environments. *Adv. Appl. Prob.* **36**, 293–317.

[7] N. Bambos, J.C. Walrand (1993). Scheduling and stability aspects of a general class of parallel scheduling systems. *Adv. Appl. Prob.* **25**, 176–202.

[8] A. Bedekar, S.C. Borst, K. Ramanan, P.A. Whiting, E.M. Yeh (1999). Downlink scheduling in CDMA data networks. In: *Proc. Globecom '99*, 2653–2657.

[9] 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.

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

[11] 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.

[12] T. Bonald, M. Jonckheere, A. Proutière (2004). Insensitive load balancing. In: *Proc. ACM Sigmetrics / Performance 2004,* 367–377.

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

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

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

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

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

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

[19] S. Das, H. Viswanathan, G. Rittenhouse (2003). Dynamic load balancing through coordinated scheduling in packet data systems. In: *Proc. Infocom 2003*.

[20] J.M. Holtzman, S. Ramakrishna (1998). A scheme for throughput maximization in a dual-class CDMA system. *IEEE J. Sel. Areas Commun.* **40**, 830–844.

[21] 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.

[22] N. Kahale, P.E. Wright (1997). Dynamic global packet routing in wireless networks. In: *Proc. Infocom '97.*

[23] 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.

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

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

[26] S. Ramakrishna (1998). *Optimal Scheduling of CDMA Systems*. Ph.D. thesis, WINLAB, Rutgers University.

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

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

[29] L. Tassiulas, A. Ephremides (1992). Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multi-hop radio networks. *IEEE Trans. Aut. Control* **37**, 1936–1938.

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

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

[32] H.Q. Ye (2003). Stability of data networks under an optimization-based bandwidth allocation. *IEEE Trans. Aut. Control* **48**, 1238–1242.