Architecture and Abstractions for Environment and Traffic Aware System-Level Coordination of Wireless Networks

Balaji Rengarajan and Gustavo de Veciana

***Abstract*—This paper presents a system level approach to interference management in an infrastructure based wireless network with full frequency reuse. The key idea is to use loose base station coordination that is tailored to the spatial load distribution and the propagation environment to exploit the diversity in a user population's sensitivity to interference. System architecture and abstractions to enable such coordination are developed for both the downlink and the uplink cases, which present differing interference characteristics. The basis for the approach is clustering and aggregation of traffic loads into classes of users with similar interference sensitivities that enable coarse grained information exchange among base stations with greatly reduced communication overheads. The paper explores ways to model and optimize the system under dynamic traffic loads where users come and go resulting in interference induced performance coupling across base stations. Based on extensive system-level simulations, we demonstrate load-dependent reductions in file transfer delay ranging from 20-80% as compared to a simple baseline not unlike systems used in the field today, while simultaneously providing more uniform coverage. Average savings in user power consumption of up to 75% is achieved. Performance results under heterogeneous spatial loads illustrate the importance of being traffic and environment aware.**

***Index Terms*—Wireless networks, cellular networks, interference management, flow level performance, performance optimization, scheduling and coordination**

I. INTRODUCTION

One way to overcome the dearth of spectrum is to consider network deployments with increased base station/access point densities. By decreasing the distance between users and their base stations, one can drastically increase capacity while reducing transmission energy requirements. Of course, this comes at a significant increase in infrastructure and management costs. There are also deleterious implications in terms of the operational regime of such networks. In particular, the proportion of users whose capacity is limited by interference from their neighbors grows. Also, as the number of base stations serving an area is increased, the coverage area and the number of users served by individual base stations decreases. This has the undesirable side effect of reducing the network's capability for statistical multiplexing and increases the burstiness' of the offered load. Thus we are faced with operating wireless systems in a highly dynamic, interference limited regime. Effectively managing inter-cell interference is essential to fully realizing the potential of broadband wireless networks, and is the focus of this paper.

Traditional approaches for mitigating interference across base stations in an infrastructure based wireless network partition resources, e.g., frequency, so that concurrent transmissions can be realized with minimal interference. Such approaches are simple and do reduce the effective interference seen by users, thus enhancing the coverage area of a base station. However, this reduction in interference is achieved at the expense of significantly diminished individual peak and overall system capacity. Reusing the entire frequency spectrum in every cell can allow us to achieve very large network capacities, provided inter-cell interference is effectively managed.

Most approaches for mitigating the effects of inter-cell interference have been studied in the context of a fixed set of users. Schemes where a centralized scheduler makes all user scheduling decisions using system wide queue and channel state information are presented in [1], [2]. Static schemes that create different reuse factors over different time periods by turning base stations on and off have been studied in [3]–[7]. A power-control based interference management scheme is proposed in [8]: users are served using one of two sets of carriers that use different power levels. A different approach that varies transmit power across time at a slow pace so as to improve performance is proposed in [9]. Fractional frequency reuse has been proposed as a mechanism for interference mitigation in OFDMA networks, allowing base stations to use different transmit powers in different frequency sub bands. Algorithms to dynamically create fractional frequency reuse patterns that maximize network utility were proposed in [10], [11]. While these algorithms respond to changes on fast time scales (seconds), they do not adapt to slow changes in the long-term load. The focus of the above schemes is to ensure that all users perceive acceptable signal to interference ratios. However, this metric does not fully describe the flow-level performance experienced by best effort users, e.g., file transfers and web browsing. The coordination scheme presented in this paper adapts to long-term (hours) spatial load variations, reducing communication overheads and directly optimizes flow-level performance.

In a realistic scenario, data requests from users are generated at random times, and the users leave when their service requirements have been met. Such dynamic systems are, in general, hard to analyze and have not been studied as extensively as their static counterparts, i.e., serving a fixed set of backlogged users. The load dynamics translate to time

This research was supported by NSF Award CNS-0721532, and an Intel Research Council Grant.

The authors are with the University of Texas at Austin. email:{balaji, gustavo}@ece.utexas.edu

varying interference that couples user capacities across base stations, and even the stability of such systems is difficult to verify, see [12]. The performance that users perceive in such dynamic systems can be very different from the performance predicted by the static model; e.g., the flow level performance of opportunistic scheduling was studied in a dynamic setting in [13], and it was demonstrated that schemes that are optimal in a static setting are sub-optimal for the dynamic setting. Potential capacity gains from inter-cell coordination in a dynamic setting were characterized in [14], and the results confirm that significant gains can be obtained through inter-cell coordination in an interference limited system.

Key idea: Base stations in wireless networks do not serve a fixed set of users concentrated at a single location, but rather serve users that tend to be distributed geographically throughout the service area. Users at diverse locations typically see very different channel gains to the neighboring base stations in the network. The key idea in this paper is to take advantage of this diversity in users' sensitivity to interference originating from the adjoining cells.

Fig. 1 illustrates this idea in the case of downlink transmissions in a two base station network. If both base stations choose to serve users near the cell edge, both users will see very high interference and low transmission rates. If Base Station B serves a nearby user (low interference sensitivity) while Base Station A serves an edge user, high data rates can be achieved to both users when B reduces its transmit power while A transmits at high power. Indeed the interference seen by the edge user would be greatly reduced, while the nearby user would suffer only a slight degradation in capacity since it has a high channel gain to its serving base station and a low channel gain to the interfering base station. The challenge lies in coordinating transmissions and choosing transmit power levels that achieve good user performance while keeping complexity and overheads low. The novelty of our work lies in the development of new abstractions, a network architecture, and associated optimizations that make such coordination practical, and efficient.

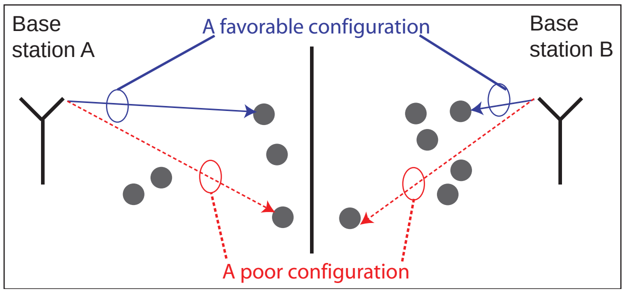


Fig. 1: Traffic abstractions to enable efficient coordination.

Contributions: In this paper, we propose a measurement-based coordination scheme that is tailored to the spatial load distribution served by the network, as well as the particular propagation environment. The proposed scheme only requires coarse grained information to be communicated among base stations over slow time scales, resulting in greatly reduced demands on the backhaul. We evaluate performance in a dynamic setting where users come and go, and the main metrics of interest are the file transfer delay or average flow throughput and the mean power expended by the transmitters. Due to space limitations, we focus solely on data traffic, yet voice and real-time traffic exhibit similar gains, albeit one has to address the fine-grained QoS requirements of such traffic. We highlight our contributions as follows:

First, we develop an approach to measure and classify a spatial population of users into a small number of user *classes* that capture average system loads, characteristics of the propagation environment, and interference sensitivities. These user classes are a critical abstraction towards reducing the complexity of the system-level optimization. To enable the optimization of *class-level* coordination schedules, one needs to properly represent the service rates that classes will see in a dynamic system. We propose an effective approximation for this which factors the intra-class variability in service capacity across users.

Second, we investigate the optimization of a coarse-grained coordination schedule. We consider various scenarios from high to low loads. Key differences arise due to the degree of dynamic interference, i.e., neighboring base stations may not always be on, and the extent to which this impacts the optimized schedule's performance may vary. We propose and evaluate various approaches to incorporate such dynamics.

Third, through extensive analysis and simulation, we illustrate the significant gains that can be achieved in terms of delay performance, power consumption at the transmitter, and substantially enhanced spatially homogeneous service to users. We further demonstrate the impact that the spatial traffic distribution can have on user performance, illustrating the importance of a scheme that is traffic aware.

The rest of this paper is organized as follows: We present our system model in Sec. II. Sec. III describes a methodology for efficiently abstracting the traffic and environment by aggregating users into representative classes. In Secs. IV-VI, we discuss methods to optimize coordinated schedules that improve user-level *downlink* performance. Sec. VII summarizes the additional benefits of base station coordination such as power savings at the base station, and increased spatial homogeneity in user performance. In Sec. VIII, we explore the benefit of being traffic-aware for heterogeneous spatial loads. Sec. IX explores the differences between the downlink and uplink scenarios, and describes an uplink coordination scheme and its performance. Finally, Sec. X concludes the paper.

II. System Model

We consider best-effort file transfers both on the uplink and downlink. User requests are assumed to arrive to the coverage area $\mathcal{X}$ as a Poisson process with a location dependent intensity $\lambda(x), x \in \mathcal{X}$. For simplicity, each user is assumed to be associated with the base station that provides the strongest signal, e.g., the geographically closest base station in the absence of shadowing. User requests arrive at random, and leave the system when the associated data transfer is completed. In the downlink scenario, we assume that base stations transmit at the specified power level when there are active associated users present, and turn off otherwise. A natural consequence of this assumption is that base stations interfere with transmissions in the neighboring cells only when serving associated users. Similarly, in the uplink scenario, no interference is generated by a cell with no active users.

In a wireless cellular network with full reuse, it is typically transmissions in the neighboring cells that generate most of the interference. In a small network, all the base stations could potentially be coordinated. Larger networks can be split into a number of independent coordinated clusters, such that the cells/sectors whose performance is tightly coupled through mutual interference are grouped together. Let N denote the number of neighboring base stations/sectors being coordinated, indexed by $b = 1, \ldots, N$. We denote by $\overline{F}_{bk}$, the mean size in bits of a file that is transferred on the downlink or the uplink. We assume that the channel between base stations and users are reciprocal. Let h_i^b denote the average channel gain between base station b and user i, and $\vec{h}_i = (h_i^b | b = 1, \ldots, N)$ represent a collection of channel gain vectors.

A. *Simulation Model*

In the sequel we will describe different methods to coordinate base station transmissions, and use extensive uplink and downlink simulations to compare their performance. In the simulations, we consider three facing sectors in a hexagonal layout of base stations with cell radius 250m - see Fig. 3a. Users associate themselves to the geographically closest base station. A carrier frequency of 1GHz, and a bandwidth of 10MHz are assumed. The maximum transmit power is limited to 10W. The base stations are assumed to be able to transmit at three different power levels: 0, 5, and 10W. Additive white Gaussian noise with power -55dBm is assumed. We consider a log distance path loss model [15], with path loss exponent 2. Shadowing, and fading are not considered in these results, but the addition of shadowing does not fundamentally change the characteristics of our measurement driven scheme, see Sec. III. File sizes are assumed to be log normally distributed, with mean 2MB. The data rate at which users are served is calculated based on the perceived SINR using Shannon's capacity with rates quantized to 0, 1, 2, 5, 10, 20, and 30Mbps. Users arrive according to a Poisson process, and except in Sec. VIII are assumed to be distributed uniformly within the simulated area. In Sec. VIII, we explore the impact of non-homogeneous user populations, and the spatial traffic model is described in detail therein. In all our simulations, mean user perceived delay is estimated within a relative error of 1%, at a confidence level of 95%.

III. Traffic Abstractions

Exploiting the diversity in user populations to mitigate the effects of interference requires base stations to adapt not only to user distributions within their cells, but also to distributions in neighboring cells. Sharing information between base stations on a per user basis would result in extremely high communication costs. So, in this section we propose to use aggregates, see Fig. 2, that allow base stations to efficiently share information about the spatial loads and sensitivities to interference.

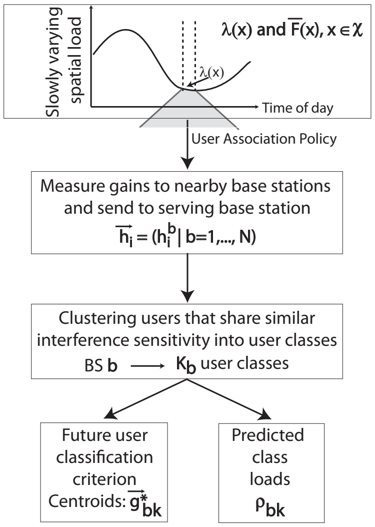
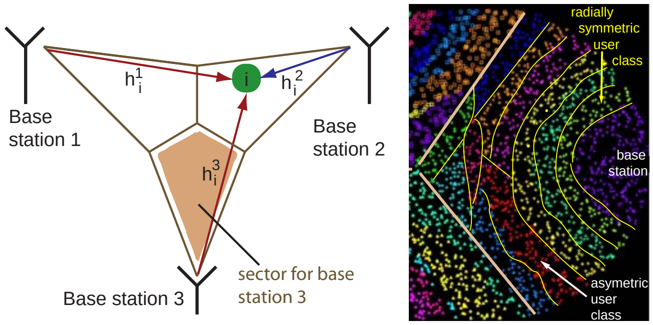

Slowly varying spatial load
λ(x) and F̄(x), x ∈ χ
λ(x)
Time of day
User Association Policy
Measure gains to nearby base stations and send to serving base station
h⃗i = (hi^b | b=1,..., N)
Clustering users that share similar interference sensitivity into user classes
BS b ⟶ Kb user classes
Future user classification criterion
Centroids: g⃗*bk
Predicted class loads
ρbk

Fig. 2: Traffic abstractions to enable efficient coordination.

In addition to short term, unpredictable variations in the load caused by individual user arrivals and departures, there are predictable long-term variations in the aggregate traffic load depending on the day of week, hour-of-day, etc. [16], [17]. Consider monitoring a user population sharing a wireless system over a long period of time, say a few hours. We shall assume that during this period, the average rate of user requests arriving at any location x remains constant, i.e., $\lambda(x)$ denotes the long-term rate of arrival of user requests at location x. Note that the traffic load might still be spatially heterogeneous. For each base station/sector b, we define K_b *user classes* that will abstract the key characteristics of the load distribution and the propagation environment. They enable base stations to measure, aggregate, and share coarse grained information about the traffic loads they support. They also drive the system-level optimization methodology described in the sequel.

User classes and class loads aggregate users (locations) that share similar sensitivity to interference from neighboring base stations. A simple way to capture these environmental conditions is to measure the average channel gains between users and neighboring base stations – this is already done in practice to facilitate handoffs. Users feedback the measured channel gain vectors $\vec{h}_i$ to their serving base stations. Fig. 3a depicts the measurements made by each user when coordinating three facing sectors in a hexagonal layout of base stations.



(a) An example scenario for coordination. (b) Example class definitions.

Fig. 3: Building user classes.

Users sharing similar gain vectors, $\vec{h}_i$, have similar susceptibility to interference from neighboring base stations on the downlink, and cause similar levels of interference at the neighboring base stations in the uplink case. Yet, in an interference limited regime, Shannon's capacity formula suggests that users transmission rates vary as the logarithm of the ratio of the received signal power to interference. Thus, for each user measurement, we define a logarithmically distorted gain vector $\vec{g}_i = (g_i^b | b = 1, \ldots, N)$, where $g_i^b = \log(h_i^b)$. In this paper, a k-means clustering algorithm [18], [19] is used to cluster measured log-gain vectors into a fixed number of user classes. Specifically, the algorithm partitions users associated with base station b into K_b clusters with centroids $\vec{g}^*_{bk}, k = 1, \ldots, K_b$, such that the mean Euclidean distance between the log-gain vectors and the centroids is minimized. Given a clustering, and the resulting centroid vectors, future users can be classified based on which centroid its log-distorted gain vector is closest to. In the sequel, we address the question of how many classes are used per base station, and the associated tradeoffs.

Fig. 3b exhibits a clustering for a sector in our example scenario where three neighboring base stations are to be coordinated. The points in the figure represent individual users, while the sets reflect their division into classes. The users near the serving base station are minimally impacted by interference, leading to radially symmetric classes that are influenced mainly by the path loss to the serving base station. Interference plays a significant part in transmissions involving users at the cell edge, and the asymmetric classes reflect the resulting discrimination based on which neighbor has the most impact. Note that in practice, due to shadowing and real environment obstructions, user classes will not result in the clean spatial partition exhibited in this example. In fact, they would instead reflect the character of the environment as well as the typical locations where the user population dwells.

With classes defined, estimating the average loads for each class under a given spatial traffic load is a simple task. Arrivals to class $k = 1, \ldots, K^b$ associated with base station/sector b are thus Poisson, with rate denoted by λ_{bk}. Define $\rho_{bk} = \lambda_{bk} \overline{F}_{bk}$ to be the mean traffic (bits per second) arriving at class k in base station b. Let $\vec{\rho} = (\rho_{bk} \,:\, b = 1, \ldots, N,\, k = 1, \ldots, K^b)$ denote the expected offered load vector. The classes may have different offered loads, capturing in part the spatial distribution of traffic supported by the system. The expected offered load vector can thus be exchanged between base stations infrequently (on the order of hours), drastically reducing communication overheads.

IV. System Abstractions For The Downlink Case

For simplicity, we will initially focus on the downlink scenario. A *joint transmission profile* represents one of the various modes in which the network can be operated. As illustrated in Fig. 4, it specifies a power profile, i.e., the transmit power level for each base station, and the associated user classes to be jointly served. The base stations are assumed to be able to transmit at one of P discrete power levels, including 0, corresponding to no transmission. The N-dimensional column vectors $\vec{p}^i$ and $\vec{c}^j$ specify the power levels and classes to be served by the base stations under power profile i and class combination j. The b^{th} component of these vectors, p_b^i and c_b^j, specify the transmit power to be used by base station b and the class to be served. The number of different power profiles is denoted by $U = P^N$, the number of class combinations by $V = \prod_{b=1}^{N} K_b$, and thus the number of joint transmission profiles is $L = UV$. Let $\mathcal{P} := \{\vec{p}^1, \ldots, \vec{p}^U\}$ and $\mathcal{C} := \{\vec{c}^1, \ldots, \vec{c}^V\}$ denote the sets of admissible joint power profiles and class combinations respectively for the N base stations, and $\mathcal{L}$ the set of joint transmission profiles. Thus, each joint transmission profile l where $l = 1, \ldots, L$ is two vectors: $\vec{p}(l) \in \mathcal{P}$ and $\vec{c}(l) \in \mathcal{C}$.

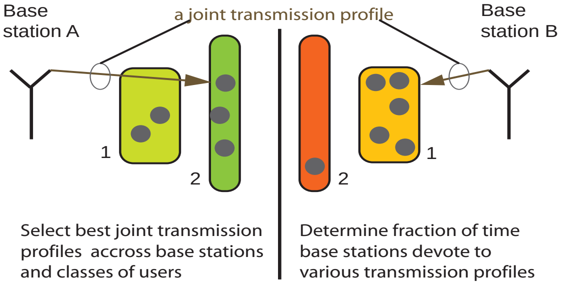

Base station A
a joint transmission profile
Base station B
1
2
2
1
Select best joint transmission profiles accross base stations and classes of users
Determine fraction of time base stations devote to various transmission profiles

Fig. 4: Illustration of a joint transmission profile.

A joint transmission schedule corresponds to the fractions of time $\vec{\alpha} = (\alpha_l \,:\, l = 1, \ldots, L)$ for which the network operates in each transmission profile. We assume that base stations are synchronized, as is the case in modern broadband systems like GSM and WiMAX, and they can (quickly) cycle through these profiles.[1] In general, this schedule will be picked to optimize a chosen performance measure, $f(\vec{\alpha})$, through an optimization of the form:

Problem 4.1: A generic optimization problem to determine a coordination schedule:

$$\min_{\vec{\alpha}} f(\vec{\alpha})$$

such that

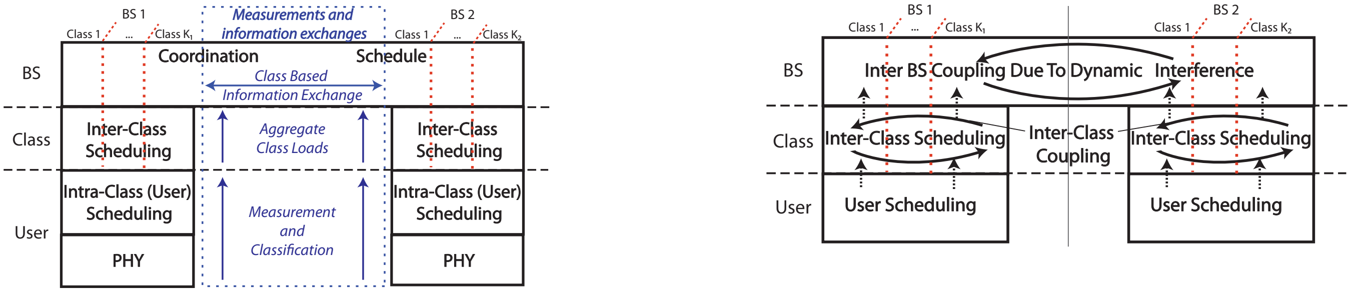
$$\rho_{bk} \quad \leq \quad R_{bk}(\vec{\alpha}),\ \forall b, k, \tag{1}$$

$$\sum_{l=1}^{L} \alpha_l \leq 1 \quad \text{and} \quad \alpha_l \geq 0,\ l = 1, \ldots, L. \tag{2}$$

Here, $R_{bk}(\vec{\alpha})$ denotes the capacity allocated to class k at base station b under schedule $\vec{\alpha}$. Eq. (1) constrains the rate allocation across classes to be one that stabilizes the network. Eqs. (2) ensure that the coordination schedule is valid.

Fig. 5a exhibits the overall system architecture used for coordination. Each user reports signal strength measurements from neighboring base stations to its serving base station. Each base station uses this information to aggregate users into classes that capture the spatial load being served, and the nature of the propagation environment. Coarse grained information about traffic loads and achievable transmission rates are exchanged between base stations at the level of user classes. The base stations can then determine the optimized

[1] Note that the resource that is subdivided for the purpose of coordination could also be frequency, or even a combination of time and frequency in an OFDMA-like system.



(a) System architecture for coordination.

(b) Modeling performance coupling.

Fig. 5: System abstractions for downlink coordination.

coordination schedule (common to all coordinating base stations), i.e., the fraction of time each transmission profile is used. The user classes are the key abstraction that allows such cross-layer, cross-base station optimizations to be carried out while keeping communication and computational overheads manageable.

Note that the transmission profiles are *not* a specification of which user to serve, only a restriction on the transmit power to be used at the base station and a recommended' class that might be beneficially served. Base stations can independently devise complimentary dynamic *inter-class* scheduling policies that serve classes other than the recommended one. Since the choice of class does not affect the interference levels observed at the neighboring cells, such inter-class scheduling does not violate the coordination schedule. Further, base stations can use any *intra-class* scheduling policy to serve users within the selected class(es). In this paper, we assume that base stations use processor sharing (or an approximation thereof) to serve the active users in the chosen class(es). Fig. 5a summarizes the relationship between the various elements of the architecture.

In order to solve Problem 4.1, we need accurate estimates for $R_{bk}(\vec{\alpha})$. However, the dynamics of the system make this a difficult task, see Fig. 5b. User performance is coupled across base stations as the capacity to users is impacted significantly by the state (transmitting or not) of neighboring base stations. A further degree of coupling, intra-base station coupling, can be introduced depending on how individual users and classes are served within each base station. If inter-class scheduling depends, for instance, on the instantaneous loads in the classes, the performance of the different classes at a base station will be coupled together. The choice of user and inter-class scheduling policy also affect the activity level of the base stations, thus impacting neighboring base stations through interference driven coupling.

Determining the exact capacity allocated by a schedule to each class when the activity levels and performance of neighboring base stations are coupled corresponds to analyzing a set of spatially coupled (through interference) queues. Systems of coupled queues have been analyzed in the past [20]–[22], but the problem is extremely difficult and closed form expressions are only available for simple scenarios with only two queues. So, for simplicity, we assume in Sec. V that the performance of the various base stations are decoupled by assuming base stations are always on; i.e., the performance (interference) seen by users does not depend on the traffic at other base stations.

As a further approximation, we study policies under which a base station is restricted to serving only the class specified by the transmission profile. If the chosen class has no active users, the base station does not opt to serve another class. We call such a policy *static scheduling*. Thus, there is no inter base station or inter class coupling. Subsequently, we evaluate the performance of a policy that allows base stations to share the excess capacity from empty scheduled classes among other associated users and thus introduces coupling among classes. In Sec. VI we will drop our assumption of decoupled base stations and present approximations for optimizing the coupled systems.

V. Optimizing The Decoupled Model

A. Static Scheduling

As the number of user classes is increased, the fidelity of the gathered information increases. However, communication overheads, and the computational complexity associated with the proposed coordination scheme also grow. Problem 4.1, for example, has a number of constraints and decision variables which respectively grow linearly and polynomially (of degree N) in the number of classes. Therefore, it is advantageous to use a relatively small number of classes. However, in this case, there may be large disparities in transmission rates of users in the same class. In order to optimize the schedule, we first need to develop good estimates for the class capacities, $R_{bk}(\vec{\alpha})$ which themselves depend on the schedule $\vec{\alpha}$. As will be seen in this section, this is not a simple matter even for a decoupled model, yet good approximations that make the optimization problem convex can be found to make this tractable.

Let the random variable I denote a randomly selected user from the system's load distribution, i.e., the distribution of user requests; thus $I = i$ corresponds to a location, and assume user i stays there until its request is completed. Let $b(i)$, and $k(i)$ be user i's base station and class respectively. Finally, let R_i^l denote the peak rate at which user i can be served under profile l, assuming all base stations are active. Note that R_i^l is zero, if a class other than $k(i)$ is served by base station $b(i)$ under profile l.

Proposition 5.1: Consider the downlink queue associated with class k at base station b. It sees an offered load of ρ_{bk} bits/sec. and a time varying capacity that depends on $\vec{\alpha}$. Suppose the rate at which base stations switch among profiles is fast compared to the time scale of the user dynamics, and the base station uses processor sharing to serve users in each class, then the queue is stable if $u_{bk} = \frac{\rho_{bk}}{R^H_{bk}(\vec{\alpha})} < 1$, where

$$R^H_{bk}(\vec{\alpha}) = \frac{1}{\mathbf{E}\left[\frac{1}{\sum_{l=1}^{L} \alpha_l R^l_I} \,\Big|\, b(I) = b,\, k(I) = k\right]}. \quad (3)$$

Further, when the queue is stable, the mean number of active users in the class is given by $\frac{u_{bk}}{1-u_{bk}}$.

Proof: If the rate at which base stations switch between the different transmission profiles is infinitely fast, the variations in rate perceived by users become negligible, and the system corresponds to a processor sharing queue operating in a ‘fluid’ regime similar to the approximation used in [23]. In this regime, a typical user I is served at the average transmission rate given by $\sum_{l=1}^{L} \alpha_l R^l_I$ if it is the only active user in the class. In this case, the time to serve user I is $\frac{\overline{F}_{bk}}{\sum_{l=1}^{L} \alpha_l R^l_I}$. The mean time to serve a user in the class is given by $\mathbf{E}\left[\frac{\overline{F}_{bk}}{\sum_{l=1}^{L} \alpha_l R^l_I}\right] = \frac{\overline{F}_{bk}}{R^H_{bk}(\vec{\alpha})}$. The total normalized load offered by the class is then given by $u_{bk} = \frac{\rho_{bk}}{R^H_{bk}(\vec{\alpha})}$. The fact that this processor sharing queue is stable when $u_{bk} < 1$ follows from the results in [13], [23], and the mean queue length of the system can be computed to be $\frac{u_{bk}}{1-u_{bk}}$ using the expression for the queue length distribution from [23]. ■

Note that $R^H_{bk}(\vec{\alpha})$ is the harmonic mean of the average transmission rates seen by the different users in class k associated with base station b. We shall refer to this as the capacity allocated to the class under schedule $\vec{\alpha}$. Unfortunately, estimating this for each $\vec{\alpha}$ requires knowledge (estimates) of the complete spatial distribution of users versus simple descriptive statistics, e.g., means and variances, and thus increased communication and computational overheads.

The arithmetic and geometric mean of the average transmission rate perceived by users are alternative estimates for the class capacity. The arithmetic mean approximation is given by:

$$\begin{aligned} R^A_{bk}(\vec{\alpha}) &= \mathbf{E}\left[\sum_{l=1}^{L} \alpha_l R^l_I \,\Big|\, b(I) = b,\, k(I) = k\right] \\ &= \sum_{l=1}^{L} \alpha_l \mathbf{E}[R^l_I \mid b(I) = b,\, k(I) = k]. \quad (4) \end{aligned}$$

The geometric mean approximation for class capacity is given by:

$$R^G_{bk}(\vec{\alpha}) = \exp(E[\log(\sum_{l=1}^{L} \alpha_l R^l_I) \mid b(I) = b,\, k(I) = k]).$$

Note that the arithmetic mean is simple to compute: it depends only on the mean rates observed by users in the class under each profile, and is linear in $\vec{\alpha}$. However, it can be shown that $R^H_{bk}(\vec{\alpha}) \leq R^G_{bk}(\vec{\alpha}) \leq R^A_{bk}(\vec{\alpha})$, whence the geometric mean is the better estimate for the harmonic mean [24]. Unfortunately, the geometric mean is also burdensome to compute, making it unsuitable.

An approximation for the geometric mean based on moments was derived in [25], and empirical studies presented in [26] show that the approximation yields accurate results. We propose using this approximation, truncated to the first and second moments, to effectively capture intra-class diversity in transmission rates. Let Σ_{bk} be the covariance matrix of the transmission rates to the users in class k in base station b, $\sigma_{bk}(l, m) = \mathbf{Cov}[R^l_I,\, R^m_I \mid b(I) = b,\, k(I) = k]$. The rate allocated to class k in base station b is approximated as

$$\begin{aligned} R^G_{bk}(\vec{\alpha}) &\approx R^A_{bk}(\vec{\alpha}) - \frac{\mathbf{Var}\left[\sum_{l=1}^{L} \alpha_l R^l_I \,\Big|\, b(I) = b,\, k(I) = k\right]}{2R^A_{bk}(\vec{\alpha})} \\ &= R^A_{bk}(\vec{\alpha}) - \frac{\vec{\alpha}^T \Sigma_{bk} \vec{\alpha}}{2R^A_{bk}(\vec{\alpha})}. \quad (5) \end{aligned}$$

Thus, the capacity allocated to all classes can be estimated with the coordinating base stations exchanging only the class means, and covariances of the transmission rates under the different profiles.

However, the estimate in Eq. (5) does not lead to constraint (1) being a provably convex function of $\vec{\alpha}$. We use the following approximation to Eq. (5) to model the allocated rates:

$$R^{GA}_{bk}(\vec{\alpha}) = R^A_{bk}(\vec{\alpha}) - \frac{\vec{\alpha}^T \Sigma_{bk} \vec{\alpha}}{c_{bk}}. \quad (6)$$

Here, $\vec{c} = (c_{bk}, b = 1, \ldots, N,\, k = 1, \ldots, K^b)$ is a positive vector that is appropriately chosen, to yield a good estimate for the class capacity.

Fact 5.1: $\left(R^A_{bk}(\vec{\alpha}) - \frac{\vec{\alpha}^T \Sigma_{bk} \vec{\alpha}}{c_{bk}}\right)^{-1}$ is a convex function of $\vec{\alpha}$, when it is positive, and $\vec{c}$ is any positive vector.

Proof: $\frac{\vec{\alpha}^T \Sigma_{bk} \vec{\alpha}}{c_{bk}}$ is convex in $\vec{\alpha}$, since the covariance matrix and thus the Hessian is positive semidefinite. Also, $-R^A_{bk}(\vec{\alpha})$ is a linear function of $\vec{\alpha}$. Thus, $-R^A_{bk}(\vec{\alpha}) + \frac{\vec{\alpha}^T \Sigma_{bk} \vec{\alpha}}{c_{bk}}$ is also convex in $\vec{\alpha}$. This implies that $R^A_{bk}(\vec{\alpha}) - \frac{\vec{\alpha}^T \Sigma_{bk} \vec{\alpha}}{c_{bk}}$ is a positive concave function. Since the reciprocal of a positive, concave function is convex, $\left(R^A_{bk}(\vec{\alpha}) - \frac{\vec{\alpha}^T \Sigma_{bk} \vec{\alpha}}{c_{bk}}\right)^{-1}$ is a convex function of $\vec{\alpha}$. ■

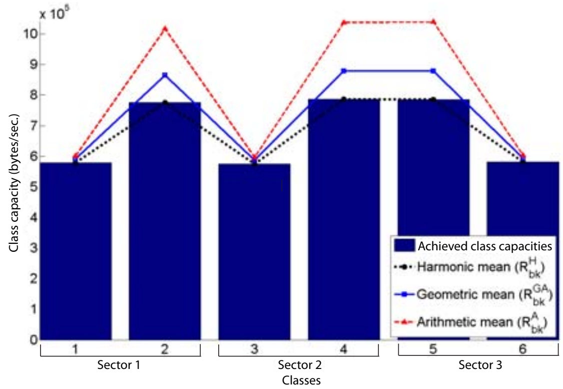


Fig. 6: Comparing the different estimates for class capacity

We examine the actual achieved class capacities, and compare it to the estimates developed above in an example scenario with three sectors and users classified into two classes per sector using the method described in Sec. III. Fig. 6 exhibits the class capacities for a fixed transmission schedule.The classification process results in classes of uneven sizes, and the classes with higher load in Fig. 6 correspond to larger fractions of the uniformly distributed users with larger intra-class variance in the rates observed by users. The schedule used allocates a larger share of the capacity to these larger classes. Both the arithmetic and geometric mean approximations are optimistic in estimating the capacity allocated to classes, but the geometric mean is much more accurate as it takes into account the variability within a class. As can be seen from the figure, this larger variance results in the arithmetic mean being too optimistic for the larger classes, and overestimates the capacity allocated to the classes by up to 20% compared to the geometric mean estimate. Our simulation results also indicate that the geometric mean approximation yields considerably better estimates for the class capacities, compared to the arithmetic mean. Next, we discuss two different strategies for optimizing user performance.

Matching Capacity and Load: The first schedule optimization approach we consider to determine the joint transmission schedule is as follows:

Problem 5.1: Determine a static, capacity maximizing, decoupled schedule based on:

$$\min_{\vec{\alpha}} \left\{ \sum_{l=1}^{L} \alpha_l \,\middle|\, \rho_{bk} \leq R_{bk}(\vec{\alpha}),\, \forall b, k, ;\, \alpha_l \geq 0,\, l = 1, \ldots, L. \right\}$$

The optimal schedule maximizes the fraction of time that the system is idle, which is a natural starting point. The optimal transmission schedule $\vec{\alpha}^*$ associated with Problem 5.1 assigns capacity to each class in proportion to the offered load. This formulation is similar to the idealized case considered in [14], and the optimal schedule stabilizes the network, if possible, for any load distribution proportional to $\vec{\rho}$ when $R_{bk}(\vec{\alpha})$ is exact, i.e., $R_{bk}(\vec{\alpha}) = R^H_{bk}(\vec{\alpha})$.

However, we propose to use the geometric approximation from Eq. (6) to estimate class capacities. To determine the constants, c_{bk}, we first solve optimization Problem 5.1 with $R_{bk}(\vec{\alpha}) = R^A_{bk}(\vec{\alpha})$, to find $\vec{\alpha}^{A*}$. We let c_{bk} be the arithmetic mean approximation of the rate allocated using schedule $\vec{\alpha}^{A*}$, $c_{bk} = R^A_{bk}(\vec{\alpha}^{A*})$. We then re-solve problem 5.1 with the geometric mean approximation.

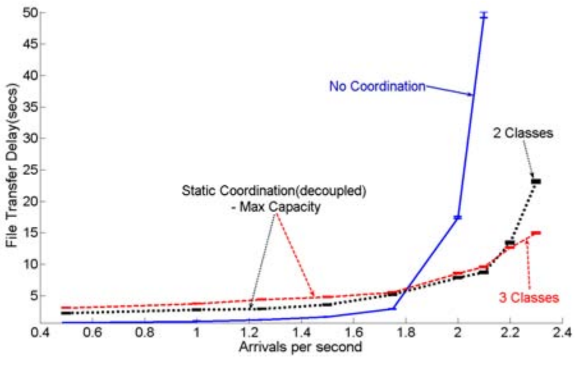


Fig. 7: Average file transfer delays under capacity maximizing static schedules.

The graph in Fig. 7 shows the average downlink file transfer delays vs. offered load obtained under three schemes: uncoordinated transmissions at the maximum power, and two static approximations with two and three user classes per base station. At higher loads, coordination performs extremely well, improving delay performance over the scheme with no coordination by over 80%. However, this is not uniformly the case, and at very low loads, the coordination scheme increases mean delays by around 50% compared to the non-coordinated scheme. Under low loads, coordinating across base stations to mitigate interference is less of a concern because the probability that neighboring base stations are simultaneously transmitting is low. Therefore, one might as well allow base stations to transmit at higher power without coordination. Also, since we are using a static schedule, the probability that there are no active users in the class scheduled at a base station is high at low loads. This leads to the base station unnecessarily wasting time while users wait their turn to get served. This is also the reason for the coordination scheme with only two classes per sector outperforming the scheme with three classes until the offered load is high enough. A larger number of classes results in base stations wasting more time when using a static schedule, as the scope for statistical multiplexing is further reduced. Splitting the load and the resources into independent small chunks results in reduced capacity for sharing, and incurrs a statistical multiplexing loss. At low loads, the gains from reduced interference levels resulting from careful coordination across base stations are not sufficient to compensate for this statistical multiplexing loss.

Delay Optimal Scheduling: When the load offered by different user classes is very different, allocating capacity proportionally to the load does not result in optimal delay performance. Classes with a larger number of users share the allocated capacity more effectively due to statistical multiplexing within the class vs. smaller' classes. Therefore, delay performance can be further improved by allocating more than a proportional share of the capacity to the smaller classes, and less to the larger classes. The following optimization minimizes the mean sum queue length across all the classes, assuming each class corresponds to a M/GI/1-PS queue, thus minimizing user-perceived delay. We continue to assume that the different base stations are decoupled.

Problem 5.2: Determine a static, delay minimizing, decoupled schedule based on:

$$\min_{\vec{\alpha}} \sum_{b=1}^{N} \sum_{k=1}^{K_b} \frac{\frac{\rho_{bk}}{R_{bk}(\vec{\alpha})}}{1 - \frac{\rho_{bk}}{R_{bk}(\vec{\alpha})}}$$

such that

$$\begin{aligned} \rho_{bk} \quad &\leq \quad R_{bk}(\vec{\alpha}),\, \forall b, k, \\ \sum_{l=1}^{L} \alpha_l \leq 1 \quad &\text{and} \quad \alpha_l \geq 0,\, l = 1, \ldots, L. \end{aligned}$$

One can show that this corresponds to a convex objective function, and the constraint set in the above optimization problem is also convex as shown above.Note that one can also consider other convex objective functions to capture other QoS metrics such as blocking rate, or other metrics such as power consumption at the base stations.

Fig. 8 exhibits the performance of the capacity maximizing schedule developed earlier vs. the above delay minimizing

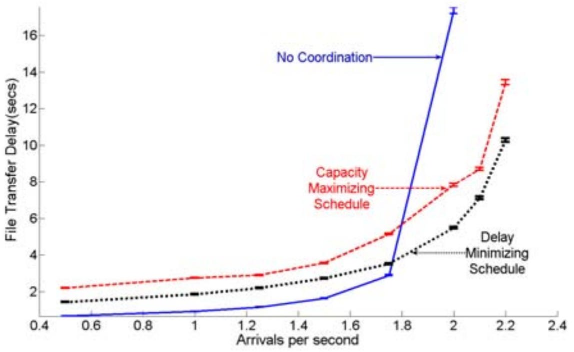


Fig. 8: Performance of capacity maximizing vs. delay optimal static, decoupled schedules with 2 classes per sector.

approach under a static schedule. Both scenarios utilize two classes per base station along with the geometric approximation in Eq. (6) to estimate the class capacities, and three transmit power levels. The queue length-minimizing approach clearly outperforms the first approach where we allocated capacity proportionally to the class loads. This is mainly because this approach takes into account the potential each class has for statistical multiplexing. We will use this queue length-minimizing approach as the basis for developing further improved joint transmission schedules in the sequel.

B. Dynamic Inter-Class Scheduling

As noted earlier, for downlink transmissions, the capacity perceived by users in neighboring base stations is independent of the user/class that a base station serves and depends only on the transmit power levels used by the various base stations. Thus, when there are no active users in the class picked by the static schedule, the base station can dynamically pick an alternate class to serve without adversely affecting any of the cooperating base stations, i.e., without increasing the interference levels perceived by users. This class can be chosen by the base station based on different criteria, such as maximizing transmission rates, or serving the class with the largest number of active users. We refer to this as inter-class scheduling.

Definition 5.1: We refer to an inter-class scheduler that upon exhausting the scheduled class performs processor sharing scheduling across all active users, as a *dynamic processor sharing inter-class scheduler*.

We found in our simulations that the delay performance of this strategy compared favorably to other policies. Note that this strategy allocates a proportionally larger rate to user classes that have a large number of active flows. When the traffic offered by all classes share similar characteristics, the optimized static schedule balances the expected number of active users in each class. Thus, this dynamic scheduling strategy attempts to align the available capacity to the particular realization of the offered load. In Fig. 9, we show results for coordination along with this type of inter class scheduling.

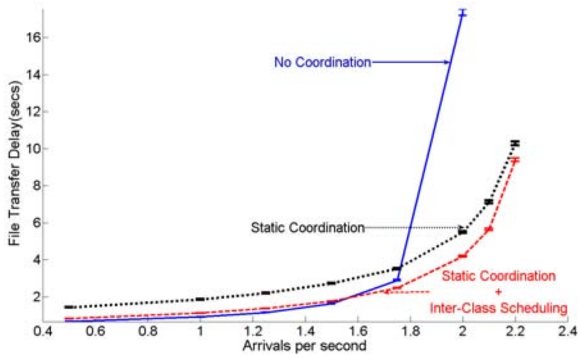


Fig. 9: Delay performance with inter-class scheduling.

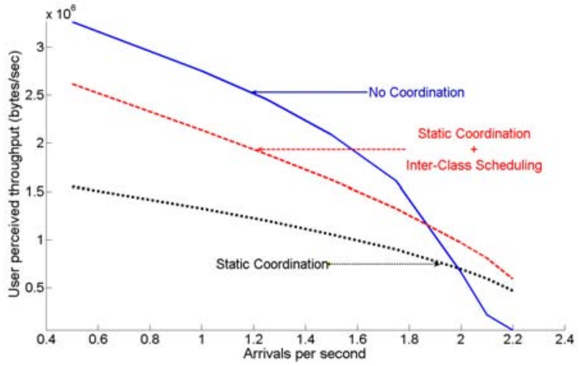


Fig. 10: User throughput with inter-class scheduling.

As can be seen in Figs. 9 and 10, inter-class scheduling significantly improves user delay performance and throughput, especially at light to moderate loads where mean delays are reduced by up to 40% as compared to the static scheme. At very low loads, it is still true that a scheme that transmits at maximum power without any coordination outperforms the coordination scheme. Attempting to coordinate transmissions at low loads results in base stations needlessly using a lower power, thus transmitting at a lower rate even when the neighboring base stations are idle. Since the probability of simultaneous transmissions occurring is minimal at low loads, coordinating is not worthwhile.

VI. Optimizing The Coupled Model

Our coordination schedules thus far have not taken into account the utilization of the neighboring base stations, and the performance coupling resulting from inter-cell interference. This is responsible for the poor performance at low loads. Determining the exact utilizations of the mutually coupled network of base stations for a particular joint transmission schedule is a difficult problem. However, if the utilizations can be estimated, the actual capacity perceived by classes in the dynamic, coupled system can be approximately determined. This would, in turn, allow us to pick better coordination schedules that explicitly take into account the degree to which the base stations are coupled.

Consider again the static coordination scheduling policies introduced in Sec. V-A. Let $\vec{u}(\vec{\alpha}) = (u_{bk}(\vec{\alpha}) : b = 1,\ldots,N,\, k = 1,\ldots,K^b)$, where $u_{bk}(\vec{\alpha})$ is the resulting utilization of class k in base station b. As the base stations switch among different transmission profiles, a base station might not transmit in a designated profile if there are no active users at that base station. As a result, users in neighboring base stations can be served at enhanced rates. This effect can be modeled as a correspondence between a profile chosen as part of the joint transmission schedule, and a number of *induced* profiles in which the network actually operates depending on class utilizations.

A base station remaining idle, with no users to serve just corresponds to using a transmit power level equal to zero, which corresponds to a valid joint transmission profile. When N base stations are being coordinated, each transmission profile can, in actual operation, result in one of up to 2^N profiles depending on which base stations are busy, or idle. Note that, these induced profiles are still a subset of $\mathcal{L}$. Let $\vec{\beta} = (\beta_m : m = 1, \ldots, L)$ be the fractions of time actually spent in each profile when the transmission schedule specified by $\vec{\alpha}$ is followed. We will approximate $\vec{\beta}$ as a function of $\vec{\alpha}$ and $\vec{u}$ as follows.

$$\beta_m(\vec{\alpha}, \vec{u}) = \sum_{l=1}^{L} \alpha_l q_l^m(\vec{u}),$$

where $q_l^m(\vec{u})$ denotes the probability that, given the transmission profile l is chosen by the schedule, the network actually operates in profile m because the corresponding set of base stations are inactive. We define the vector $\vec{s}^{lm} = (s_b^{lm} : b = 1, \ldots, N)$ that takes binary values as follows: $s_b^{lm} = 1$ if $p_b(l) = p_b(m)$, and 0 otherwise. We estimate $q_l^m(\vec{u})$ assuming that the busy periods of the queues corresponding to the classes in different base stations are independent, i.e.,

$$q_l^m(\vec{u}) = \begin{cases} 0 & \text{if } \vec{c}(l) \neq \vec{c}(m), \\ 0 & \text{if } \vec{p}(m).(\vec{p}(l) - \vec{p}(m)) \neq 0, \\ \prod_{b=1}^{N} (u_{bc_b(l)})^{s_b^{lm}} (1 - u_{bc_b(l)})^{(1-s_b^{lm})} & \text{otherwise.} \end{cases}$$

Note that the network can only operate in a transmission profile m that allocates the same transmit power level as l, (or zero) to the base stations. This is captured by the second case in the equation above. The fraction of time actually spent by the network in each induced profile can be computed in a similar fashion in the case of the dynamic coordination policy, except that q_l^m depends on the probability that there are no active users in any of the classes associated with a base station.

We propose to compute a joint transmission schedule optimizing users' delay performance while taking into account the coupling across base stations iteratively. Let u_{bk}^z, R_{bk}^z represent the utilization, and rate estimates for the classes used in iteration z. Here, $\vec{\beta}^z = (\beta_m^z : m = 1, \ldots, L)$ denotes the computed resultant schedule induced by the choice of time fractions $\vec{\alpha}^z = (\alpha_l^z : l = 1, \ldots, L)$ in iteration z, and is a function of u_{bk}^z, and $\vec{\alpha}^z$. We let $\vec{\alpha}^{z*}$ denote the optimal coordination schedule found in iteration z, and $\vec{\beta}^{z*}$ the resultant induced schedule. Initially, $u_{bk}^1 = 1, \forall b, k$, and $R_{bk}^1 = R_{bk}^{A^*}$, and

$$\begin{aligned} \beta_m^z(\vec{\alpha}^z, \vec{u}^z) &= \sum_{l=1}^{L} \alpha_l^z q_l^m(\vec{u}^z) \\ u_{bk}^{z+1} &= \frac{\rho_{bk}}{R_{bk}^{(z)}(\vec{\beta}^{(z)*})}, \forall b, k. \end{aligned}$$

The optimization problem solved at each iteration is:

Problem 6.1: Determining a delay minimizing schedule for the coupled network:

$$\min_{\vec{\alpha}^z} \sum_{b=1}^{N} \sum_{k=1}^{K_b} \frac{\frac{\rho_{bk}}{R_{bk}^z(\vec{\beta}^z)}}{1 - \frac{\rho_{bk}}{R_{bk}^z(\vec{\beta}^z)}}$$

such that

$$\begin{aligned} \rho_{bk} &\leq R_{bk}^z(\vec{\beta}^z), \forall b, k, \\ \sum_{l=1}^{L} \alpha_l^z &\leq 1, \\ \alpha_l^z &\geq 0, l = 1, \ldots, L. \end{aligned}$$

In the simulations that follow, we use the following geometric rate approximation based on Eq. (6):

$$R_{bk}^z(\vec{\beta}^z) = R_{bk}^{GA}(\vec{\beta}^z) = R_{bk}^A(\vec{\beta}^z) - \frac{\vec{\beta^z}^T \Sigma_{bk} \vec{\beta}^z}{2R_{bk}^{(z-1)}(\vec{\beta}^{(z-1)*})}$$

The objective function, and constraints in Problem 6.1 are convex, since $\vec{\beta}^z$ is a linear function of $\vec{\alpha}$, and the composition of a convex function and an affine function preserves convexity. This ensures that the problem can be efficiently solved at each iteration.

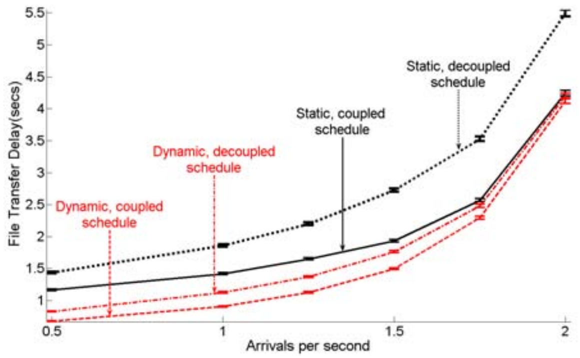


Fig. 11: Average file transfer delays under schedules factoring inter-base station coupling, with 2 classes per sector.

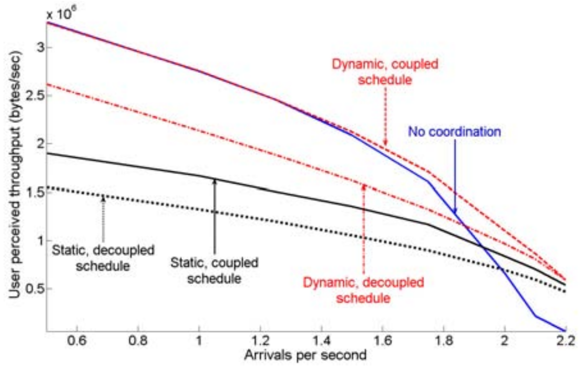
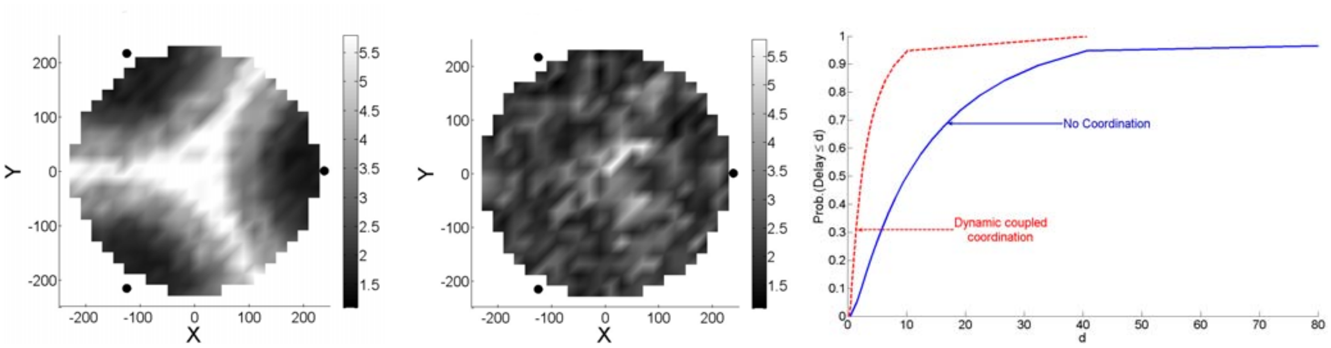


Fig. 12: Average user throughput under schedules factoring inter-base station coupling, with 2 classes per sector.

Fig. 11 illustrates the reduction in the average user-perceived delays that is achieved using two iterations in the above formulation. Here, we do not show the delay performance of the scheme with no coordination for clarity. Fig. 12 shows the increased user throughputs achieved from



(a) Spatial delay: No coordination (b) Spatial delay: Dynamic coordination (c) CDF of user delay

Fig. 13: Distribution of user-perceived delay

this coordination scheme, and also compares against the non-coordinated case. Now, at low loads, the coordinated transmission schedule does not penalize performance by restricting the transmit power level used by the base stations. The coordinated schedule performs as well as random scheduling at very low loads, when the probability of simultaneous transmissions at neighboring base stations is extremely low. At moderate to high loads, an optimized coordinated scheduling scheme factoring the effect of coupling across base stations considerably outperforms the non-coordinated network, decreasing mean delays by over 80% as compared to a non-coordinated scheme. This ensures that the coordination scheme achieves good delay performance irrespective of the load on the network.

VII. Power Savings and Spatial Homogeneity

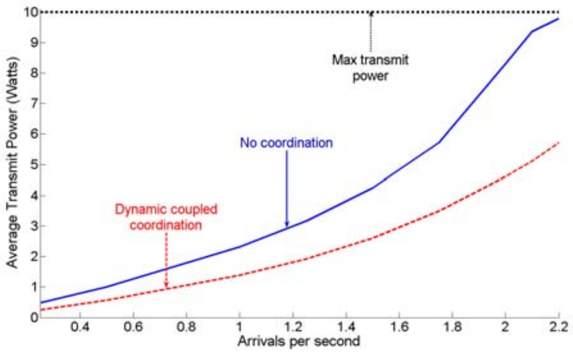


Fig. 14: Average power consumed at the base stations.

In addition to improving delay performance and capacity, coordination has further benefits. As shown in Fig. 14, the average power expended by the base station is substantially reduced when coordination is used, e.g., 45% when the arrival rate is 2 users per second. This suggests a reduction in cooling costs at the base station, and also indicates that we can further improve delay performance if the base stations were subject to mean power constraints, and could transmit at higher peak power levels.

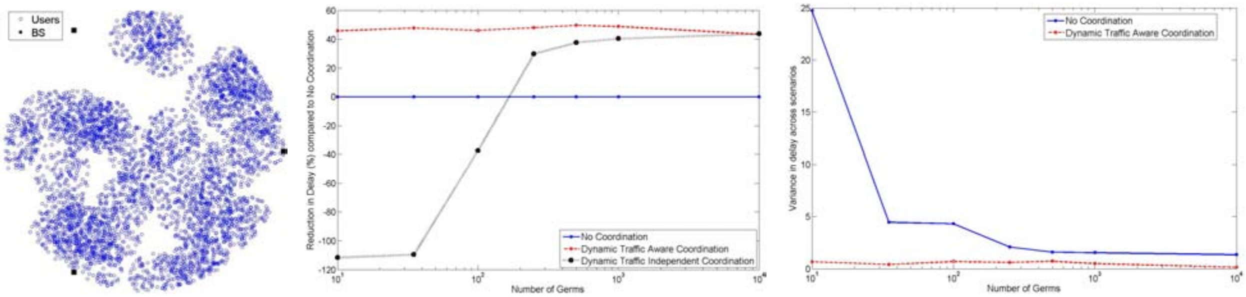
Figs. 13a, and 13b show the spatial delay distribution induced when no coordination is used, and the coordination scheme that minimizes the overall queue length, with $\lambda = 1.75$. As shown in Fig. 13b, when coordination is used, the average delays seen by users at different locations are much more spatially homogeneous. In particular, with no coordination users at the edge experience very poor performance. Under coordination, users' experience is virtually decoupled from their location in the coverage area.

Fig. 13c exhibits the distribution of delay across all users, when $\lambda = 2$. Coordination improves delay performance for all users, not just the ones at the edge. This is because the coordination scheme increases the probability that there are no active users at a base station. Thus, even though users close to the base stations are potentially served using lower transmit power levels, they benefit from lower interference levels.

VIII. The Importance of being Traffic Aware

In a real-world wireless network, the traffic load is unlikely to be spatially homogeneous and may exhibit significant variations over time. For example, at different times of the day, one might see concentrations in different regions, e.g., coffee houses, lunch spots, public transportation, or depending on congestion patterns, etc. We shall explore the potential gains from coordination in such scenarios. In particular, we are interested in understanding the degree to which optimizing for a particular load is beneficial. For example, if a fixed interference mitigation scheme such as a static fractional frequency reuse pattern is used, a natural choice is to optimize for a uniform distribution of users. We shall evaluate the performance of our dynamic coordination scheme when it is optimized for a uniform load, versus when it is tuned to the particular spatial traffic load.

Our clustered traffic model is as follows. User locations are constrained to a subset of the simulated area determined by the realization of a Boolean germ-grain model [27]. The grains of the Boolean model are discs of fixed radius, while the germs are distributed uniformly within the simulated area. The probability that an arrival's location falls in any of the discs is equal. The density of users at various points within the cell depends on the number of grains covering it. The density of users in areas covered by multiple grains is high, resembling a hotspot. Fig. 15a exhibits a realization of the spatial load with 70 germs, and discs with radius equal to one fifth the radius of the cell are used. Note that there are regions within the cell with sparse user densities, and others where users tend to cluster. As the number of germs increases, the arrivals process converges to a homogeneous Poisson process. A small number



(a) A clustered user population

(b) Reduction in average file transfer delays

(c) Variance in average file transfer delay

Fig. 15: A scenario with non-homogeneous spatial load

of germs represents a user population that is highly clustered, with large variations in user densities within the coverage area.

In our simulations, the number of germs is varied from 10 to 10,000 to simulate various degrees of clustering in the spatial load. For each case, we investigate the performance in twenty different realizations for the Boolean model. As explained previously, the actual load on the system is highly dependent on the spatial characteristics of the traffic. To roughly evaluate performance under vastly different spatial loads, we normalize the overall arrival rate so that the actual loads are comparable. Specifically, we choose the arrival rate that results in the base stations being 95% utilized, assuming all base stations transmit at maximum power all the time even if they have nothing to send. This operating point is computed using the harmonic mean, as described in Sec. V.

Fig. 15b depicts the reduction in delay achieved by the two schemes compared to the non-coordinated case. It is clear that when the actual traffic being served is highly clustered (small number of germs), the traffic-independent coordination scheme performs much worse. In fact the average delays experienced by users are more than doubled vs the case with no coordination. As the number of germs is increased, and the spatial distribution of users approaches the uniform distribution, the traffic-independent scheme performs better than the non-coordinated one, and eventually catches up to the traffic-aware scheme. The reduction in delay achieved by the traffic aware scheme appears independent of clustering in the loads. Note however, that our normalization is imperfect, and in fact the measured loads were lower for scenarios subject to clustered loads. Thus we conjecture that subject to the same system load the gain achieved by the traffic aware will increase if the spatial load exhibits higher random clustering.

Fig. 15c exhibits the variance across the scenarios under the traffic-aware scheme and the case where no coordination is used. This variance is induced by the sensitivity to intercell interference, and because different locations are affected very differently by interference. A non-coordinated system that serves a varying, non-homogeneous spatial distribution of users is prone to excessive variations in user perceived performance, and can experience very poor delay performance during time periods when it has to support a user population that is "poorly situated". The traffic aware coordination scheme is successful in shielding users from varying spatial loads, and achieves relatively homogeneous performance independent of where a user population lies. This decoupling of performance from both the variable spatial distribution of load, and the location of the users is a significant benefit.

IX. System Abstractions For The Uplink Case

The impact of interference on the uplink is less pronounced than on the downlink. The key difference between the downlink and uplink cases results from the shift in the source of interference. On the downlink, the interference perceived by a user is independent of the particular user that is scheduled at the neighboring base stations.On the uplink, users' transmissions create interference at the neighboring base stations and a change in the location of the user can drastically alter the resulting degree of interference. This automatically modulates the interference caused at a base station as users at different locations are scheduled.

Compared to the downlink case where edge users always see high interference from active neighbors, the number of scenarios where users are severely limited by interference on the uplink is reduced. Consider the scenario depicted in Fig. 16 when all transmissions are at full power. In the downlink case shown in Fig. 16a, the edge user receives interference that is very close to the strength of the received signal. However, the interference at BS B is very low on the uplink as shown in Fig. 16b and the edge user's rate is impacted much less by interference. BS A does perceive higher interference in the uplink scenario than the near user does on the downlink. However, the interference is still much weaker than the received signal at BS A, and the nearby user can be served at high rates.

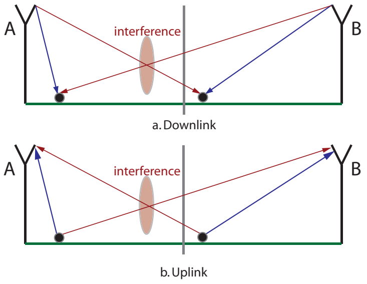


Fig. 16: Difference between uplink and downlink scenarios.

Users close to their serving base station typically have high channel gains to the serving base station and low channel gains to the neighboring base stations. Such users cause very low interference at their neighbors, and due to high received signal strengths are not severely affected by interference themselves. Users at the cell edge cause very high interference at the neighboring base stations, and additionally have to cope with low channel gains to the serving base station. Thus, the strength of interference seen by a base station depends both on the transmit power chosen by the users transmitting in the neighboring cells as well as the channel from the interfering user to the base station. While knowledge of the transmit power level at the neighboring base stations was sufficient to predict the interference received by a user or user class on the downlink, a different coordination mechanism is required for uplink coordination.

The abstractions used for uplink coordination are similar to the downlink case. The difference is that *power profiles* are replaced by *joint interference profiles*. As noted earlier, fixing the transmit power used in all the cooperating cells/sectors is not sufficient to predict the interference at the base stations. The interference profiles directly bound the average interference that each cell is allowed to cause on neighboring base stations, so now an uplink transmission profile is specified by the combination of an interference profile and a class vector.

The maximum interference caused by transmissions at a cell to a neighboring base station is set at one of Q discrete levels, including 0. The $N \times N$ matrix $\vec{\mathbf{q}}^i$ specifies bounds on the interference each sector can cause at each of its neighboring base stations under interference profile i. The maximum average interference that transmissions at sector b can cause at sector m under interference profile i is denoted $q^i_{b,m}$. Note that $q^i_{b,b} = \infty$, for $b = 1, \ldots, N$. The number of different interference profiles is denoted by $U' = Q^{(N(N-1))}$, and the number of joint transmission profiles is $L' = U'V$. Let $\mathcal{Q} := \{\vec{\mathbf{q}}^1, \ldots, \vec{\mathbf{q}}^{U'}\}$ and $\mathcal{C} := \{\vec{c}^1, \ldots, \vec{c}^V\}$ denote the sets of admissible joint interference profiles and class combinations respectively for the N base stations. Thus, each joint transmission profile l where $l = 1, \ldots, L'$ is given by: $\tilde{\mathbf{q}}(l) = \tilde{\mathbf{q}}^i \in \mathcal{Q}$ and $\vec{c}(l) = \vec{c}^j \in \mathcal{C}$. A joint uplink transmission schedule corresponds to the fractions of time $\vec{\alpha} = (\alpha_l \colon\ l = 1, \ldots, L')$ for which the network uses each transmission profile.

A. Determining user power levels under interference profiles

Consider a user u served by base station/sector b'. Recall that $\vec{h}_u$ is the channel gain vector corresponding to user u. Users choose the largest transmit power that ensures that the average interference caused at all the neighboring base stations meets the constraints. The transmit power chosen by user u under interference profile i is given by $\min_{b=1,\ldots,n} \frac{q^i_{b',b}}{h^b_u}$.

B. Estimating class rates

Each user can calculate the minimum rate achieved under each transmission profile after calculating the transmit power and using the upper bounds on received interference specified in the corresponding interference profile. Thus, the harmonic mean of the user rates provides a lower bound on the effective rate at which users in any class are served. Any of the estimates presented in Sec. V can then be used as an approximation of the class rates under a particular coordination schedule. In the simulation results presented in the sequel, we use the geometric mean rate approximation.

C. Optimizing the schedule

In order to optimize the user perceived delay performance, we use the methodology described in Sec. VI, with the transmission profiles defined as the combination of an interference profile and a class vector. The optimization problem solved at iteration z is:

Problem 9.1: Determine a delay minimizing schedule for the coupled network based on:

$$\min_{\vec{\alpha}^z} \sum_{b=1}^{N} \sum_{k=1}^{K_b} \frac{\frac{\rho_{bk}}{R^z_{bk}(\vec{\beta}^z)}}{1 - \frac{\rho_{bk}}{R^z_{bk}(\vec{\beta}^z)}}$$

such that

$$\begin{aligned} \rho_{bk} \quad &\leq \quad R^z_{bk}(\vec{\beta}^z),\ \forall b, k, \\ \sum_{l=1}^{L'} \alpha^z_l \leq 1 \quad &\text{and} \quad \alpha^z_l \geq 0,\ l = 1, \ldots, L'. \end{aligned}$$

While optimizing user performance is valuable, reducing the energy consumption while maintaining acceptable user performance is likely to be an important concern on the uplink. The arithmetic mean of the users' transmit powers under each transmission profile provides an approximation of the power consumed. Let $J_b(l)$ denote the average power consumed at base station b under transmission profile l. The average power consumption under a joint transmission schedule $\vec{\alpha}$ can then be estimated as $\sum_{l=1}^{L'} \sum_{b=1}^{N} \alpha^z_l J_b(l)$. The coordination framework presented above can be modified to minimize the weighted sum of the mean user delay and the mean power consumption under a coordination schedule. The weight chosen, denoted by γ, represents the relative importance of conserving energy versus minimizing delay, and is a parameter that can be adjusted. The objective function to be minimized at each iteration in the methodology of Sec. VI is given by

$$\sum_{b=1}^{N} \sum_{k=1}^{K_b} \frac{\frac{\rho_{bk}}{R^z_{bk}(\vec{\beta}^z)}}{1 - \frac{\rho_{bk}}{R^z_{bk}(\vec{\beta}^z)}} + \gamma \sum_{l=1}^{L'} \sum_{b=1}^{N} \alpha^z_l J_b(l).$$

In the sequel, we again use processor sharing as the intra-class scheduling policy and dynamic processor sharing as the inter-class scheduling discipline - see Def. 5.1.

D. Uplink Performance

We simulate uplink file transfers in the three sector scenario shown in Fig. 3a, with user requests distributed as a homogeneous Poisson process in space. Fig. 17 exhibits plots of the mean delay performance, while Fig. 18 shows plots of the average throughput achieved under the delay-minimizing joint uplink transmission schedule. As a result of using a methodology that accounts for inter-base station coupling, we see that the delay and throughput performance always equals or improves on those obtained for the uncoordinated scheme. At high loads, mean delay is improved by about 40% when 2 classes are used per base station, and by up to 80% when

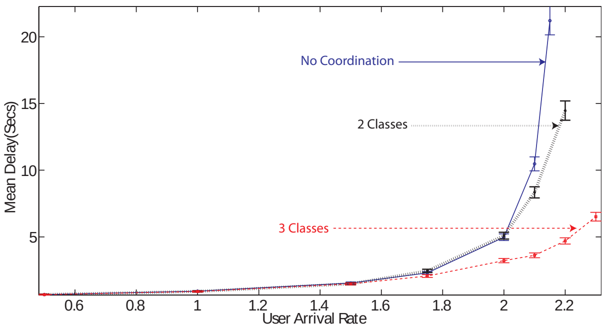


Fig. 17: Uplink delay performance under schedules factoring inter-base station coupling.

3 classes are used per base station. The average throughput is increased by up to 27% when 2 classes are used per sector and by up to 90% when 3 classes are used. At moderate to high loads, using a finer grain classification of users results in significant performance gains. Since individual users adjust their transmit powers to satisfy the constraints imposed by the interference profile, the variability in rates across users in an interference profile is increased. Users near the boundary have low channel gains to their serving base station, and are additionally forced to use lower power levels in order to limit the interference that they cause. Using a larger number of classes improves the estimates for the class rates, and also enables the schedule to accurately differentiate between users at different locations.

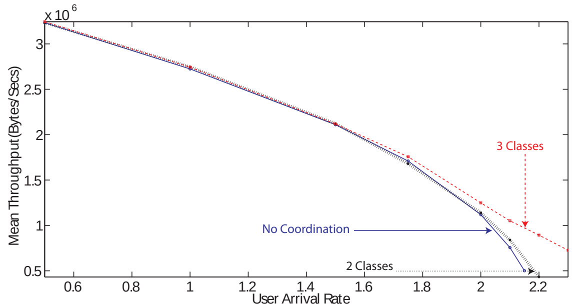


Fig. 18: Average throughput under schedules factoring inter-base station coupling.

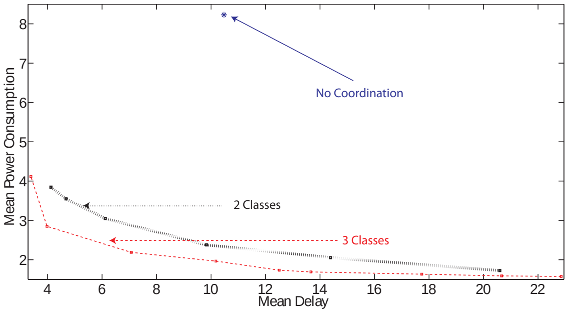


Fig. 19: Power-Delay tradeoffs on the uplink.

Fig. 19 exhibits plots of the mean delay achieved against the average power consumption under the non-coordinated system as well as the coordinated schedule that minimizes a weighted sum of mean delay and mean power. The overall rate at which users arrive into the system is fixed at 2.1 arrivals per second and they are assumed to be spatially homogeneous. The weight γ is varied to demonstrate the trade-offs between energy saving and performance that can be achieved through coordination. Average energy consumption can be decreased by up to 75% through coordination while achieving delay performance identical to the non-coordinated system. Note that even when the coordination scheme is tuned to minimize user perceived delay performance, the average power consumption is lowered by approximately 50% relative to the non-coordinated system. This improvement in performance and energy efficiency is achieved while simultaneously ensuring that the average delays seen by users at different locations are much more spatially homogeneous relative to the case with no coordination, similar to the downlink case (graph omitted).

X. Conclusion and Future Work

We have proposed a low complexity, system-level approach that substantially improves performance perceived by best-effort users without requiring high channel measurement and estimation, communication, and computational overheads. The proposed approach simultaneously achieved spatially homogeneous performance while also reducing the transmit power requirements. While future wireless networks could include physical layer techniques such as interference cancellation, such techniques are likely to be imperfect due to associated measurement and estimation errors. As demonstrated in [28], significant gains can be achieved through overlaying a system level approach like the one proposed in this paper. System-level coordination can also be profitably used in the case of (packet) delay sensitive traffic, as long as suitable complementary dynamic user scheduling schemes are developed to meet users' QoS requirements. A factor that we have not considered in this paper is user mobility. Mobile users simply transition from one class to another as they move about within the network, and can potentially be treated as premature departures from a class arriving at another. These topics are left for future research.

References

[1] N. Kahale and P. E. Wright, "Dynamic global packet routing in wireless networks," in *IEEE INFOCOM*, vol. 3, Apr. 1997, pp. 1414–1421.

[2] S. Das, H. Viswanathan, and G. Rittenhouse, "Dynamic load balancing through coordinated scheduling in packet data systems," in *IEEE INFOCOM*, vol. 1, 2003, pp. 786–796.

[3] T. K. Fong *et al.*, "Radio resource allocation in fixed broadband wireless networks," *IEEE Trans. Commun.*, vol. 46, no. 6, pp. 806–818, Jun. 1998.

[4] K. K. Leung and A. Srivastava, "Dynamic allocation of downlink and uplink resource for broadband services in fixed wireless networks," *IEEE J. Select. Areas Commun.*, vol. 17, no. 5, pp. 990–1006, May 1999.

[5] X. Qiu and K. Chawla, "Resource assignment in a fixed broadband wireless system," *IEEE Communications Letters*, vol. 1, no. 4, pp. 108–110, Jul. 1997.

[6] A. Ghasemi and E. S. Sousa, "Distributed intercell coordination through time reuse partitioning in downlink CDMA," in *IEEE Wireless Communications and Networking Conference*, vol. 4, Mar. 2004, pp. 1992–1997.

[7] K. Chawla and X. Qiu, "Quasi-static resource allocation with interference avoidance for fixed wireless systems," *IEEE J. Select. Areas Commun.*, vol. 17, no. 3, pp. 493–504, Mar. 1999.

[8] J. Li *et al.*, "A static power control scheme for wireless cellular networks," in *IEEE INFOCOM*, vol. 2, 1999, pp. 932–939.

[9] X. Wu, A. Das, J. Li, and R. Laroia, "Fractional power reuse in cellular networks," in *Proceedings of the 44th Allerton Conference on Communication, Control, and Computing*, September 2006.

[10] A. L. Stolyar and H. Viswanathan, “Self-organizing dynamic fractional frequency reuse for best-effort traffic through distributed inter-cell coordination,” in *Infocom*, 2009.

[11] K. Son *et al.*, “Dynamic association for load balancing and interference avoidance in multi-cell networks,” in *WiOpt*, April 2007, pp. 1–10.

[12] S. Borst, M. Jonckheere, and L. Leskela, “Stability of parallel queueing systems with coupled service rates,” *Discrete Event Dynamic Systems*, vol. 18, no. 4, pp. 447–472, 2008.

[13] S. Borst, “User-level performance of channel-aware scheduling in wireless data networks,” in *INFOCOM 2003*, vol. 1, March-April 2003, pp. 321 – 331.

[14] T. Bonald, S. Borst, and A. Proutiere, “Inter-cell scheduling in wireless data networks,” in *European Wireless Conference*, 2005.

[15] T. S. Rappaport, *Wireless Communications: Principles and Practice,2/E*. Prentice Hall PTR, 2002.

[16] S. Borst *et al.*, “Dynamic optimization in future cellular networks,” *Bell Labs Technical Journal*, vol. 10, no. 2, pp. 99–119, 2005.

[17] S. Borst, I. Saniee, and A. Whiting, “Distributed dynamic load balancing in wireless networks,” in *International Teletraffic Congress*, 2007, pp. 1024–1037.

[18] M. Anderberg, *Cluster Analysis for Applications*. Academic Press, 1973.

[19] A. Jain and R. Dubes, *Algorithms for Clustering Data*. Prentice Hall, 1988.

[20] S. Borst, O. Boxma, and P. Jelenkovic, “Coupled processors with regularly varying service times,” in *IEEE INFOCOM 2000*, vol. 1, 2000, p. 157164.

[21] S. Borst, O. Boxma, and M. van Uitert, “The asymptotic workload behavior of two coupled queues,” *Queueing Systems*, vol. 43, no. 1-2, pp. 81–102, January 2003.

[22] G. Fayolle and R. Lasnogorodski, “Two coupled processors: The reduction to a Riemann–Hilbert problem,” *Wahrscheinlichkeitstheorie*, no. 3, pp. 1–27, Jan. 1979.

[23] T. Bonald, S. Borst, and A. Proutiere, “How mobility impacts the ow-level performance of wireless data systems,” in *INFOCOM 2004. Twenty-third AnnualJoint Conference of the IEEE Computer and Communications Societies*, vol. 3, 2004, pp. 1872–1881 vol.3.

[24] G. Hardy, J. E. Littlewood, and G. Polya, *Inequalities*. Cambridge, 1997.

[25] W. E. Young and R. H. Trent, “Geometric mean approximations of individual security and portfolio performance,” *The Journal of Financial and Quantitative Analysis*, vol. 4, no. 2, pp. 179–199, Jun. 1969.

[26] W. H. Jean and B. P. Helms, “Geometric mean approximations,” *The Journal of Financial and Quantitative Analysis*, vol. 18, no. 3, pp. 287–293, Sep. 1983.

[27] D. Stoyan, W. Kendall, and J. Mecke, *Stochastic Geometry and its Applications*. J. Wiley & Sons, Chichester, 1995.

[28] B. Rengarajan, “Self organizing networks: building traffic and environment aware systems,” Ph.D. dissertation, The University of Texas at Austin, 2009 (in preparation).

Balaji Rengarajan is a doctoral candidate in the Electrical and Computer Engineering department at the University of Texas at Austin. He joined the University of Texas in Fall ’02, and received his M.S. in Electrical Engineering in May 2004. He received his B.E. in Electronics and Communication from the University of Madras in May 2002. He was the recipient of a 2003 Texas Telecommunications Engineering Consortium (TxTEC) graduate fellowship.

Gustavo de Veciana (S’88-M’94-SM’01-F’09) received his B.S., M.S, and Ph.D. in electrical engineering from the University of California at Berkeley in 1987, 1990, and 1993 respectively. He is currently a Professor at the Department of Electrical and Computer Engineering at the University of Texas at Austin. He served as the Associate Director and then Director of the Wireless Networking and Communications Group (WNCG) 2004-2008. His research focuses on the design, analysis and control of telecommunication networks. Current interests include: measurement, modeling and performance evaluation; wireless and sensor networks; architectures and algorithms to design reliable computing and network systems. Dr. de Veciana has served as editor for the IEEE/ACM Transactions on Networking, and as co-chair of ACM CoNEXT 2008. He is the recipient of General Motors Foundation Centennial Fellowship in Electrical Engineering, an NSF Foundation CAREER Award 1996, co-recipient of the IEEE William McCalla Best ICCAD Paper Award 2000, and co-recipient of the Best Paper in ACM Transactions on Design Automation of Electronic Systems, 2002-2004.