Architecture and Abstractions for Environment and Traffic Aware System-Level Coordination of Wireless Networks: The Downlink Case

Balaji Rengarajan
Dept. of ECE, UT Austin

Gustavo de Veciana
Dept. of ECE, UT Austin

***Abstract*—Two ways to substantially enhance wireless broadband capacity are full frequency reuse and smaller cells, both of which result in operational regimes that are highly dynamic and interference limited. This paper presents a system-level approach to interference management, that has reasonable backhaul communication and computation requirements. The basis for the approach is clustering and aggregation of measurements of the spatial diversity in sensitivity to interference associated with average user populations. This enables the system to exchange information and optimize coordinated transmission schedules using only coarse grained data. The paper explores various ways of optimizing such schedules: from a static, decoupled version to a dynamic version capturing user-level scheduling, fluctuating loads and inter-cell interference that couples base stations' performance. Based on extensive system-level simulations, we demonstrate reductions in file transfer delay ranging from 20–80%, from light to heavy loads, as compared to a simple baseline not unlike those in the field today. This improvement is achieved while providing more uniform coverage, and reducing base station power consumption by up to 45%.**

I. Introduction

One way to overcome a 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 a cellular 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 static user population. Centralized joint user scheduling schemes, requiring large amounts of information to be conveyed to a centralized scheduler, are presented in [1], [2]. The centralized scheduler also has to solve a highly complex optimization problem based on the queue and channel states of all the users in the network to make scheduling decisions. Alternatively, static schemes using different reuse factors over different time periods to protect vulnerable users have been considered, see e.g., [3]–[6]. A quasi-static scheme based on a similar principle is presented in [7]. The above schemes only considered base stations that either transmit at maximum power, or are turned off. They also do not take into consideration the impact of using adaptive modulation and coding schemes. 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]. The users then track the varying channel conditions and this information is used by the base station to effectively schedule transmissions.

The focus of these schemes is to ensure that all users perceive acceptable signal to interference ratios. However, this metric does not fully describe the performance experienced by best effort users. Further, the characteristics of the user population being served do not influence the power control policy, leaving scope for further improvement. In a realistic scenario, data requests from users are generated at random times, and the users leave when their service requirements have been met. This dynamic system is, in general, very hard to analyze and has not been studied as extensively as the static version, i.e., serving a fixed set of backlogged users. The actual performance that users perceive in the dynamic system 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 [10], and it was demonstrated that schemes that are optimal in a static setting are sub-optimal for the dynamic setting. Such load dynamics also translate to time varying interference seen by users, and further impact the performance of schemes designed to mitigate inter-cell interference.

Potential capacity gains from inter-cell coordination in a dynamic setting were characterized in [11], and the results confirm that significant gains can be obtained through inter-cell coordination in an interference limited system. For a practical system, the delay performance experienced by users at typical system loads is an important consideration. The static capacity-optimal schedule developed in [11] is not a practicable solution for a system at light to moderate loads. Also, the system model considered in [11] is idealized, and would in reality be prohibitively complex in terms of the communication, and computation overhead required.

Contributions: In this paper, we propose a measurement-based 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, and 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 metric of interest is file transfer delay or average throughput. 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.

The key idea is to take advantage of the diversity in users' sensitivity to interference originating from the adjoining cells – this is not new. The novelty of our work lies in the development of new abstractions, a network architecture, and associated optimizations that make this practical, and efficient. Our focus is on coordination to improve downlink performance – a subsequent work will address the quite different uplink case. 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 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. 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.

The rest of this paper is organized as follows: We sum up the system model in Section II. Section III describes the methodology for efficiently abstracting the traffic and environment through aggregating users into representative classes. In Sections IV-VI, we discuss methods to determine coordinated schedules that improve user-level performance, in order of increasing effectiveness. Section VII summarizes the additional benefits of base station coordination such as power savings at the base station, and increased spatial homogeneity in user performance. Finally, Section VIII concludes the paper.

II. System Model

In a wireless cellular network, 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 $b = 1, \dots, N$. User requests arrive at random, and leave the system when the associated data transfer on the downlink is completed. For simplicity, each user is assumed to be served by a single fixed base station. We let $\vec{h}_i = (h_i^b | b = 1, \dots, N)$ be a collection of channel gain vectors, where h_i^b is the gain from base station b to user i, and is measured by each user and fed back to the serving base station. Fig. 1 depicts the measurements made by each user when coordinating three facing sectors in a hexagonal layout of base stations. This is the canonical example we will consider throughout this paper.

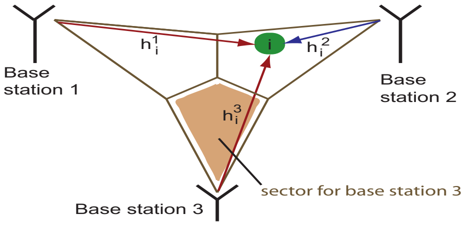


Fig. 1. An example scenario for coordination.

A. Traffic Model

User requests are assumed to arrive to the network as a Poisson process with rate λ. For each base station/sector b, we define K_b *user classes* that are used to abstract key characteristics of the load distribution and the propagation environment. Each user request is classified into a user class. Arrivals to class $k = 1, \dots, K^b$ associated with base station/sector b are thus Poisson, with rate denoted by λ_{bk}. Base stations have a file to transmit to each associated user, with mean file size $\overline{F}_{bk}$ bits. 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 $\vec{\rho} = (\rho_{bk} : b = 1, \dots, N, k = 1, \dots, K^b)$ denote the expected offered load vector. Fig. 2 illustrates a scenario with two base stations, and two classes per base station. The classes may have different offered loads, capturing in part the spatial distribution of traffic supported by the system.

B. Service Model

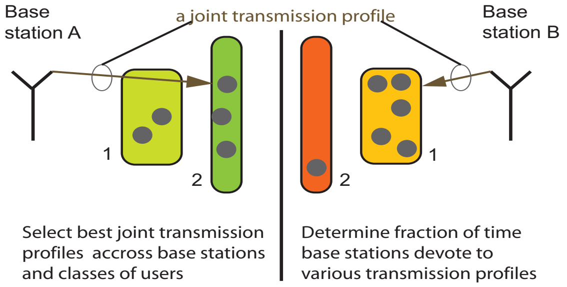

Base station A
a joint transmission profile
Base station B
1
2
2
1
Select best joint transmission profiles accross base stations and classes of users
Determine fraction of time base stations devote to various transmission profiles

Fig. 2. Illustration of a joint transmission profile.

A *joint transmission profile* represents one of the various modes in which the network can be operated. As illustrated in Fig. 2, it specifies a power profile, i.e., the transmit power level for each base station, and the associated user classes to be jointly served. Note that this is *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 complementary dynamic user/packet scheduling policies to serve their users. For simplicity, in this paper, we assume that base stations use processor sharing scheduling (or an approximation thereof) to serve the active users in a class.

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. Thus, each joint transmission profile l where $l = 1, \ldots, L$ is two vectors: $\vec{p}(l) = \vec{p}^i \in \mathcal{P}$ and $\vec{c}(l) = \vec{c}^j \in \mathcal{C}$.

A joint transmission schedule corresponds to the fractions of time $\vec{\alpha} = (\alpha_l : l = 1, \ldots, L)$ for which the network uses each transmission profile. In general, this schedule will be picked to optimize a chosen performance measure, $f(\vec{\alpha})$, through an optimization of the form:

Problem 2.1: A generic optimization problem to determine a coordination schedule:

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

such that

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

$$\sum_{l=1}^{L} \alpha_l \leq 1, \tag{2}$$

$$\alpha_l \geq 0, l = 1, \ldots, L. \tag{3}$$

Here, $R_{bk}(\vec{\alpha})$ denotes the capacity allocated to class k at base station b by the schedule $\vec{\alpha}$. Eq. (1) constrains the rate allocation across classes to be one that stabilizes the network. Eqs. (2), and (3) ensure that the coordination schedule picked is a valid one. In the sequel, we will describe different methods to determine joint transmission schedules, and use extensive simulations to compare their performance. The following section describes the simulation model in detail.

C. Simulation Model

In the simulations, we consider three facing sectors in a hexagonal layout of base stations with cell radius 250m. 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 restricted 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 [12], with path loss exponent 2. Shadowing, and fading are not considered in these preliminary results, but the addition of shadowing does not fundamentally change the characteristics of our measurement driven scheme, as noted in Sec. III-A. Users arrive according to a Poisson process, and are distributed uniformly within the simulated area. 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. The mean user perceived delay is estimated within a relative error of 1%, at a confidence level of 95%.

III. Abstracting the Traffic and the Environment

User classes and class loads aggregate users (locations) that share similar sensitivity to interference from neighboring base stations. They enable base stations to measure, aggregate, and share coarse grained information about the traffic loads they support. They also drive our system-level optimization, e.g., Problem 2.1, which has a number of constraints and decision variables which respectively grow linearly and polynomially (of degree N) in the number of classes. 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 grows. Therefore, it is advantageous to use a relatively small number of classes. However, in this case, there may be large disparities in transmission rates among users in the same class. In order to solve Problem 2.1, one must properly capture the capacities $R_{bk}(\vec{\alpha})$ that are allocated to user classes under different schedules. As will be seen in this section, this is not a simple problem, yet good approximations that make the optimization problem convex can be found to make this tractable.

A. Aggregation of Users into Classes

Consider monitoring a user population sharing a wireless system during a period of time. As shown in Fig. 1, a simple way to capture the 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 sharing similar gain vectors, $\vec{h}_i$, have similar susceptibility to interference from neighboring base stations. 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 measured user, let us define a logarithmically distorted gain vector $\vec{g}_i = (g_i^b | b = 1, \ldots, N)$, where $g_i^b = \log(h_i^b)$. Users sharing similar log-gain vectors $\vec{g}_i$ will share similar transmission rates under the various power profiles. In this paper, a k-means clustering algorithm [13], [14] 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. With classes defined, estimating the average loads for each class under a given spatial traffic load is a simple task.

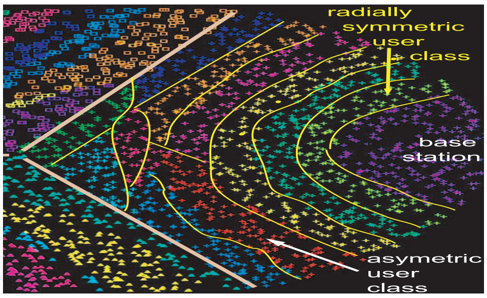


Fig. 3. An example of class definitions.

Fig. 3 exhibits a clustering for a sector in our example scenario where three neighboring base stations are to be coordinated. Note that in practice, due to shadowing and real environment obstructions, user classes will not result in the 'smooth' structure or spatial locality exhibited in this example. In fact, they would instead reflect the character of the environment as well as the typical locations where a user population tends to dwell.

B. Estimating Class Rates

Let the random variable I denote a randomly selected user from the system's load distribution, i.e., $I = i$ corresponds to a location, and assume user i stays there until his request is completed. Let $b(i)$, and $k(i)$ be user i's base station and class respectively. Finally, let R_i^l be the maximum 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 3.1: Consider the downlink queue associated with class k at base station b. It sees an offered load of ρ_{bk} bits/sec., and 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 $u_{bk} = \frac{\rho_{bk}}{R_{bk}^H(\vec{\alpha})} \leq 1$, where

$$R_{bk}^H(\vec{\alpha}) = \frac{1}{\mathbf{E}\left[\frac{1}{\sum_{l=1}^L \alpha_l R_I^l} \,\middle|\, b(I) = b,\, k(I) = k\right]}. \qquad (4)$$

Further, when the queue is stable, the mean number of active users associated with 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 [15]. In this regime, a typical user I is served at the average transmission rate given by $\sum_{l=1}^L \alpha_l R_I^l$ if it is the only active user in the class. In this case, the time to serve user I is $\frac{\overline{F}_{bk}}{\sum_{l=1}^L \alpha_l R_I^l}$. The mean time to serve a user in the class is given by $\mathbf{E}\left[\frac{\overline{F}_{bk}}{\sum_{l=1}^L \alpha_l R_I^l}\right] = \frac{\overline{F}_{bk}}{R_{bk}^H(\vec{\alpha})}$. The total normalized load offered by the class is then given by $u_{bk} = \frac{\rho_{bk}}{R_{bk}^H(\vec{\alpha})}$. The fact that this processor sharing queue is stable when $u_{bk} \leq 1$ follows from the results in [10], [15], and the mean queue length of the system can be computed to be $\frac{u_{bk}}{1-u_b k}$ using the expression for the queue length distribution from [15]. ■

Note that $R_{bk}^H(\vec{\alpha})$ is the harmonic mean of the average transmission rates seen by the different users in class k in base station b. We denote it the capacity allocated to the class under schedule $\vec{\alpha}$. Unfortunately, estimating this for each $\vec{\alpha}$ requires knowledge of the complete distribution of users versus simple descriptive statistics, e.g., means and variances, which would reduce both communication and computational overheads.

The arithmetic and geometric mean of the average transmission rate perceived by users are two alternatives to estimate class capacity. The arithmetic mean approximation is given by:

$$\begin{aligned} R_{bk}^A(\vec{\alpha}) &= \mathbf{E}\left[\sum_{l=1}^L \alpha_l R_I^l \,\middle|\, b(I) = b,\, k(I) = k\right] \\ &= \sum_{l=1}^L \alpha_l \mathbf{E}[R_I^l \mid b(I) = b,\, k(I) = k]. \qquad (5) \end{aligned}$$

The geometric mean approximation for class capacity is given by:

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

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 [16]. Unfortunately, the geometric mean is also burdensome to compute, making it unsuitable.

An approximation for the geometric mean based on moments was derived in [17], and empirical studies presented in [18] 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 X_{bk} be the covariance matrix of the transmission rates to the users in class k in base station b, $X_{bk}(l,m) = \mathbf{Cov}[R^l_I,\, R^m_I \mid b(I) = b,\, k(I) = k]$. The rate allocated to class k in base station b is approximated as

$$\begin{aligned} R^G_{bk}(\vec{\alpha}) &\approx R^A_{bk}(\vec{\alpha}) - \frac{\mathbf{Var}\left[\sum_{l=1}^{L} \alpha_l R^l_I \,\middle|\, b(I) = b,\, k(I) = k\right]}{2R^A_{bk}(\vec{\alpha})} \\ &= R^A_{bk}(\vec{\alpha}) - \frac{\vec{\alpha}^T X_{bk} \vec{\alpha}}{2R^A_{bk}(\vec{\alpha})}. \end{aligned} \tag{6}$$

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.

Our simulation results indicate that the geometric mean approximation yields considerably better estimates for the class capacities, compared to the arithmetic mean. However, the estimate in Eq. (6) does not lead to constraint (1) being a provably convex function of $\vec{\alpha}$. We use the following approximation to Eq. (6) to model the allocated rates:

$$R^{GA}_{bk}(\vec{\alpha}) = R^A_{bk}(\vec{\alpha}) - \frac{\vec{\alpha}^T X_{bk} \vec{\alpha}}{c_{bk}}. \tag{7}$$

Here, $\vec{c}$ is a positive constant that is appropriately chosen, to yield a good estimate for the class capacity.

Fact 3.1: $\left(R^A_{bk}(\vec{\alpha}) - \frac{\vec{\alpha}^T X_{bk}\vec{\alpha}}{c_{bk}}\right)^{-1}$ is a convex function of $\vec{\alpha}$, when it is positive, and c is any positive constant.

Proof: $\frac{\vec{\alpha}^T X_{bk}\vec{\alpha}}{c_{bk}}$ 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, $-R^A_{bk}(\vec{\alpha}) + \frac{\vec{\alpha}^T X_{bk}\vec{\alpha}}{c_{bk}}$ is also convex in $\vec{\alpha}$. This implies that $-\left(-R^A_{bk}(\vec{\alpha}) + \frac{\vec{\alpha}^T X_{bk}\vec{\alpha}}{c_{bk}}\right)$ is a positive concave function. Since the reciprocal of a positive, concave function is convex, $\left(R^A_{bk}(\vec{\alpha}) - \frac{\vec{\alpha}^T X_{bk}\vec{\alpha}}{c_{bk}}\right)^{-1}$ is a convex function of $\vec{\alpha}$. ■

IV. Static Scheduling

The key element of base station coordination for downlink transmission is the joint selection of a coordinated schedule. Determining the exact capacity allocated by a schedule to each class in the coupled system corresponds to analyzing a set of spatially coupled (through interference) queues. Systems of coupled queues have been analyzed in the past [19]–[21], but the problem is extremely difficult and closed form expressions are available only in the case of simple scenarios with two coupled queues. For the moment, we assume that the performance of the various base stations are decoupled, and base stations always have users to serve. We might think of this as a heavily loaded, or saturated regime. We then check if our assumption of decoupling leads to a reasonable allocation of resources.

A. Matching Capacity and Load

The first approach we consider is to frame the following optimization problem to determine the joint transmission schedule:

Problem 4.1: Determining a static, capacity maximizing, decoupled schedule:

$$\min_{\vec{\alpha}} \sum_{l=1}^{L} \alpha_l$$

such that

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

The optimal schedule maximizes the fraction of time that the system is idle, which is a natural starting point. The optimal transmission schedule $\vec{\alpha}^*$ assigns capacity to each class in proportion to the offered load. This formulation is similar to the one in [11], 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., $R_{bk}(\vec{\alpha}) = R^H_{bk}(\vec{\alpha})$.

We use the geometric approximation from Eq. (7) to estimate the class capacities. To determine the constants, c_{bk}, we first solve optimization Problem 4.1 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 $\vec{\alpha}^{A*}$, $c_{bk} = R^{A*}_{bk} = R^A_{bk}(\vec{\alpha}^{A*})$.

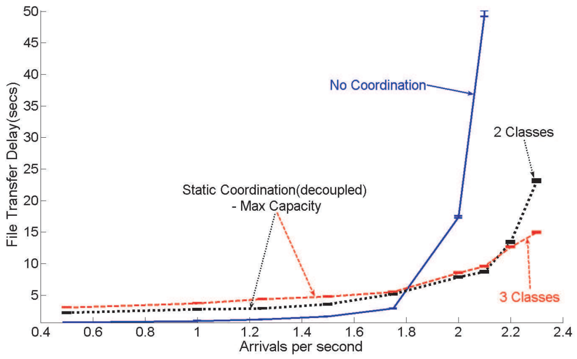


Fig. 4. Average file transfer delays under capacity maximizing static schedules.

The graph in Fig. 4 shows average downlink file transfer delays vs. offered load 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 using two classes 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.

B. 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. We continue to assume that the different base stations are decoupled. The following optimization minimizes the sum queue length across all the classes, assuming each class corresponds to a M/GI/1-PS queue, thus minimizing user-perceived delay.

Problem 4.2: Determining a static, delay minimizing, decoupled schedule:

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

such that

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

The constraint set in the above optimization problem is convex, as shown in Sec. III-B.

Fact 4.1: $\sum_{b=1}^{N} \sum_{k=1}^{K_b} \frac{\frac{\rho_{bk}}{R_{bk}(\vec{\alpha})}}{1-\frac{\rho_{bk}}{R_{bk}(\vec{\alpha})}}$ is a convex function of $\vec{\alpha}$, if $\frac{\rho_{bk}}{R_{bk}(\vec{\alpha})}$ is convex.

Proof: Let $u_{bk}(\vec{\alpha}) = \frac{\rho_{bk}}{R_{bk}(\vec{\alpha})}$. Then, $\frac{\frac{\rho_{bk}}{R_{bk}(\vec{\alpha})}}{1-\frac{\rho_{bk}}{R_{bk}(\vec{\alpha})}} = \frac{u_{bk}(\vec{\alpha})}{1-u_{bk}(\vec{\alpha})}$. $\frac{u_{bk}(\vec{\alpha})}{1-u_{bk}(\vec{\alpha})}$ is a convex non-decreasing function of u_{bk}, and $u_{bk}(\vec{\alpha})$ is a convex function of $\vec{\alpha}$. Since the composition of a convex, non-decreasing function and a convex function is convex, $\frac{u_{bk}(\vec{\alpha})}{1-u_{bk}(\vec{\alpha})}$ is a convex function of $\vec{\alpha}$. Therefore, the sum $\sum_{b=1}^{N} \sum_{k=1}^{K_b} \frac{u_{bk}(\vec{\alpha})}{1-u_{bk}(\vec{\alpha})}$ is also convex. ■

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.

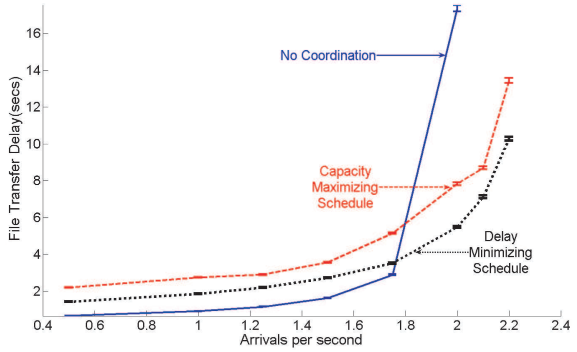


Fig. 5. Comparing the performance of capacity maximizing and delay optimal static, decoupled schedules with 2 classes per sector.

The performance of the capacity maximizing schedule developed in Sec. IV-A is compared to the above formulation which minimizes the overall queue length under a static schedule. Both scenarios utilize two classes per base station, and three transmit power levels. The queue length-minimizing approach clearly outperforms the first heuristic that 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 schedules in the sequel.

V. Dynamic Inter-Class Scheduling

Note that, in 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.

The dynamic scheduling strategy that we adopted is to serve all active users associated with a base station according to a processor sharing mechanism when the scheduled class has no active users. 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 those classes that have a large number of active users. 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 instantiation of the offered load. In Fig. 6, we show results for coordination along with dynamic inter class scheduling.

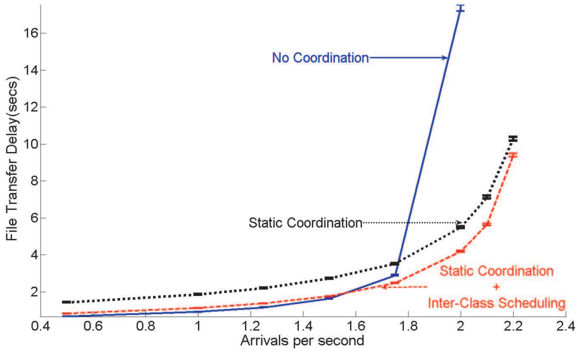


Fig. 6. Average file transfer delays under delay-minimizing static, decoupled schedules complemented by dynamic inter class scheduling with 2 classes per sector.

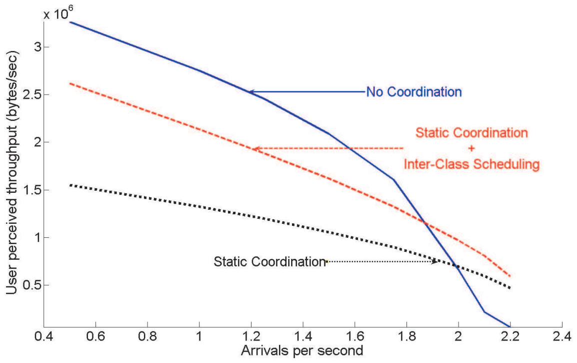


Fig. 7. Average user throughput under delay-minimizing static, decoupled schedules complemented by dynamic inter class scheduling with 2 classes per sector.

As can be seen in Figs. 6 and 7, complementary dynamic 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 System

Our coordination schedules thus far have not taken into account the utilization of the neighboring base stations, and the coupling induced by 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. IV. Let $\vec{u}(\vec{\alpha}) = (u_{bk}(\vec{\alpha}) : b = 1,\dots,N, k = 1,\dots,K^b)$, 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 is a valid choice. 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 $\vec{\beta} = (\beta_m : m = 1,\dots,L)$ be the fractions of time actually spent in each profile when the transmission schedule specified by $\vec{\alpha}$ is followed.

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

Here, $q_l^m(\vec{u})$ is the probability that the network happens to operate in profile m, based on the states of the base station queues, when transmission profile l is the one chosen by the transmission schedule. We define the vector $\vec{s}^{lm} = (s_b^{lm} : b = 1,\dots,N)$ 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 assuming that the busy periods of the queues corresponding to the classes in different base stations are independent, i.e.,

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

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. $\vec{\beta}^z = (\beta_m^z : m = 1,\dots,L)$ denotes the computed resultant schedule induced by the choice of time fractions $\vec{\alpha}^z = (\alpha_l^z : l = 1,\dots,L)$ in iteration z, and is a function of u_{bk}^z, and $\vec{\alpha}^z$. $\vec{\alpha}^{z*}$ denotes the optimal coordination schedule found in iteration z, and $\vec{\beta}^{z*}$ the resultant induced schedule. Initially, $u^1_{bk} = 1$, $\forall b, k$, and $R^1_{bk} = R^{A^*}_{bk}$, and

$$\begin{aligned} \beta^z_m(\vec{\alpha}^z, \vec{u}^z) &= \sum_{l=1}^{L} \alpha^z_l q^m_l(\vec{u}^z) \\ u^{z+1}_{bk} &= \frac{\rho_{bk}}{R^{(z)}_{bk}(\vec{\beta}^{(z)*})}, \forall b, k. \end{aligned}$$

The optimization problem solved at each iteration is:

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

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

such that

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

In the simulations that follow, we use the following geometric rate approximation based on Eq. (7):

$$R^z_{bk}(\vec{\beta}^z) = R^{GA}_{bk}(\vec{\beta}^z) = R^A_{bk}(\vec{\beta}^z) - \frac{\vec{\beta^z}^T X_{bk} \vec{\beta}^z}{2R^{(z-1)}_{bk}(\vec{\beta}^{(z-1)*})}$$

The objective function, and constraints in optimization 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.

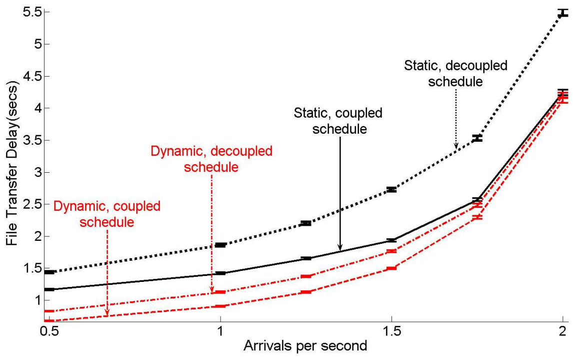


Fig. 8. Average file transfer delays under delay-minimizing schedules that account for inter-base station coupling, with 2 classes per sector.

Fig. 8 illustrates the reduction in 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. 9 shows the increased user throughputs achieved by 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, the coordinated scheduling scheme that factors in the effect of coupling across base stations considerably outperforms the non-coordinated network, decreasing mean delays by over 80% compared to a non-coordinated scheme. This ensures that the coordination scheme achieves good delay performance irrespective of the load on the network.

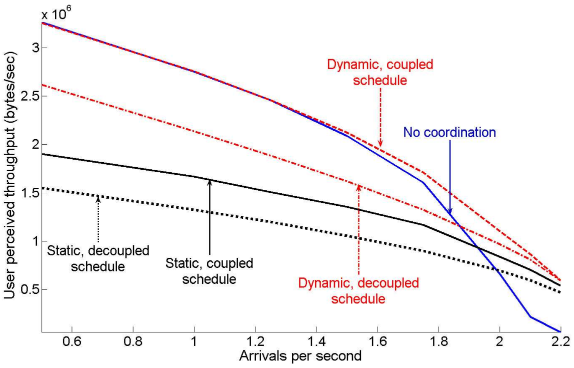


Fig. 9. Average user throughput under delay-minimizing schedules that account for inter-base station coupling, with 2 classes per sector.

VII. Power Savings and Spatial Homogeneity

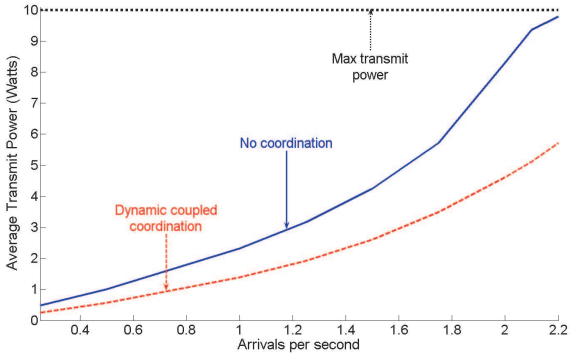
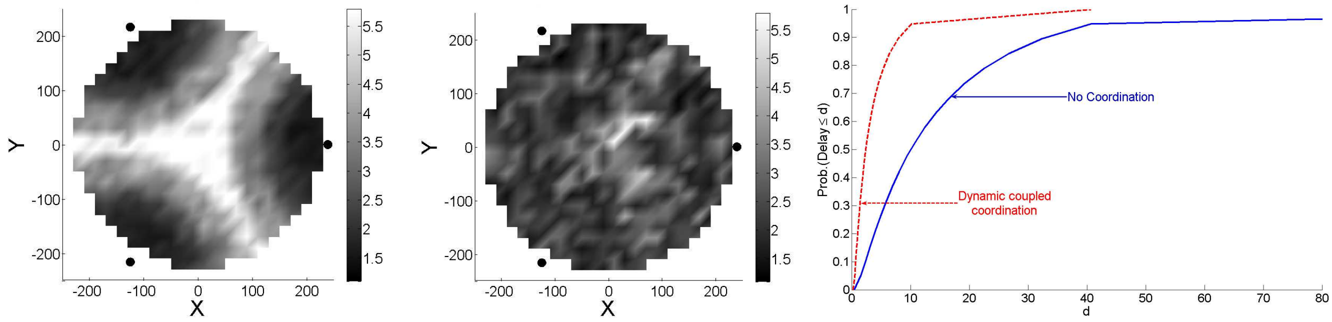


Fig. 10. Average power consumed at the base stations.

In addition to improving delay performance and capacity, coordination has further benefits. As shown in Fig. 10, 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 are allowed to transmit at higher peak power levels.

Fig. 11a, and 11b shows the spatial delay distribution induced by the scheme without coordination, and the coordination scheme that minimizes the overall queue length, with $\lambda = 1.75$. As shown in Fig. 11b, when coordination is used, the average delays seen by users at different locations are much more spatially homogeneous relative to the case with no coordination. In particular, with no coordination users at the edge experience very poor performance. Under coordination,



(a) Spatial delay: No coordination

(b) Spatial delay: Dynamic coordination

(c) CDF of user delay

Fig. 11. Distribution of user-perceived delay

users' experience is virtually decoupled from their location in the coverage area.

Fig. 11c plots 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. Conclusion and Future Work

We focused on a low complexity, system-level approach that improves performance perceived by best-effort users on the downlink 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. 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. The proposed coordination scheme can also be extended to improve uplink performance. However, the interference levels perceived by the receiving base station in uplink transmissions depends both on the power levels used in the neighboring cells, as well as the positions of the interfering users. Therefore, complementary dynamic scheduling schemes need to be carefully designed for the uplink to extract the maximum possible gains from coordination. 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. In the future, we intend to pursue these topics, and improve and extend the proposed system level approach.

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, P. S. Henry, K. K. Leung, X. Qiu, and N. K. Shankaranarayanan, "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, N. B. Shroff, and E. K. P. Chong, "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] S. Borst, "User-level performance of channel-aware scheduling in wireless data networks," in *INFOCOM 2003*, vol. 1, March-April 2003, pp. 321 – 331.

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

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

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

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

[15] T. Bonald, S. Borst, and A. Proutiere, "How mobility impacts the flow-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.

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

[17] 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.

[18] 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.

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

[20] 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.

[21] G. Fayolle and R. Lasnogorodski, "Two coupled processors: The reduction to a riemann–hilbert problem," *Wahrscheinlichkeitstheorie*, no. 3, pp. 1–27, Jan. 1979.