Practical adaptive user association policies for wireless systems with dynamic interference

Balaji Rengarajan* and Gustavo de Veciana†
*IMDEA Networks, Madrid, Spain. Email: balaji.rengarajan@gmail.com
†Dept. of ECE, University of Texas at Austin. Email: gustavo@ece.utexas.edu

***Abstract*—We study the impact of user association policies on flow-level performance in interference-limited wireless networks. Most research in this area has used static interference models (neighboring base stations are always active) and resorted to intuitive objectives such as load balancing. In this paper, we show that this can be counterproductive in the presence of dynamic interference which couples the transmission rates to users at various base stations. We propose a methodology to optimize the performance of a class of coupled systems, and apply it to study the user association problem. We show that by properly inducing load asymmetries, substantial performance gains can be achieved relative to a load balancing policy (e.g., 15 times reduction in mean delay). We present a practical, measurement-based, interference-aware association policy that infers the degree of interference-induced coupling and adapts to it. Systematic simulations establish that both our optimized static and adaptive association policies substantially outperform various dynamic policies which can, in extreme cases even be susceptible to Braess's paradox like phenomena, i.e., an increase in the number of base stations can lead to worse performance under greedy association policies. Further, these results are robust to changes in file size distributions, large-scale propagation parameters, and spatial load distributions.**

I. Introduction

The high demand for wireless capacity and the increasing volume of traffic mandates the efficient use of available radio resources. Wireless capacity can be substantially enhanced by reusing the entire frequency spectrum at every transmitter instead of sacrificing individual peak and overall system capacity by partitioning it. This increased system capacity and spectral efficiency is achieved at the expense of increased interference. Even in the case of WLANs with frequency reuse, high densities of users in large scale networks could lead to high interference due to the limited number of orthogonal frequencies available under the present standards.

The bursty nature of traffic in typical wireless systems results in dynamic interference which couples performance in the system in a complex manner. For such coupled systems, stability is fairly difficult to establish, and performance is particularly hard to optimize. The capacity of such a system as well as the actual performance that users perceive can be very different from that predicted by a saturated model that assumes that transmitters are always on, see for example [1]–[3]. Without having access to good performance models, many researchers have resorted to intuitive objectives such as load balancing across system resources. In this paper, we show that such load balancing, be it greedily done by users or across the system, may be counter productive when there is dynamic coupling due to interference.

This research was supported by NSF Award CNS-0917067, an Intel Research Council Grant, and IMDEA Networks. A preliminary version of this paper was presented at NET-COOP 2009.

Let us consider some examples where dynamic coupling impacts network functions. Consider the user association problem exhibited in Fig. 1a. Assume that the base stations share the same spectrum, so they interfere with each other when they are concurrently active, which in turn reduces their transmission capacity to users. For simplicity, assume user requests to download files arrive uniformly between base stations 1 and 2. A basic problem in such networks is to decide which base station should serve a new user request. If both the network and traffic demands are *symmetric*, one might intuitively expect that a static policy that associates arrivals with the closest base station, i.e., the one that delivers the strongest signal, and thus balances the offered load would be 'optimal'. Surprisingly, we will see that this is not the case.

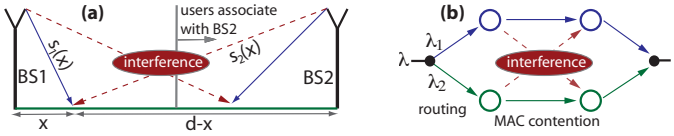


Fig. 1: User association (a) and MAC/routing functions (b) are subject to dynamic coupling.

A second example is exhibited in Fig. 1b where wireless nodes relay traffic. Assume nodes contend at random for a shared channel. Depending on the amount of traffic and interference they see, one might optimize nodes' contention probability for the channel so as to minimize overall packet delays. Clearly, performance here is a complex function of the dynamic traffic loads, contention probabilities, and interference seen by nodes. The third example, also shown in the same figure, concerns routing traffic across paths that are link/node disjoint. Unfortunately, transmissions along the paths may directly (or even indirectly) interfere with each other. Should a packet flow with rate λ be split across the two paths, or is it better to route traffic on a single path?

The above exemplify the relevance of dynamic coupling in optimizing network functions at various layers. In the above cases, assuming symmetry in loads and/or the network, one might imagine load balancing might be a good objective, but this need not be the case. For example, it may be preferable to route traffic on a single path so as to avoid interference

across paths. In this paper, we focus on the terminal association problem which, as we will see, is already fairly complex. As mentioned earlier, when the channels and traffic load are symmetric, one might expect that associating users' requests with the closest base station might be a good strategy. This corresponds to splitting the load evenly between the two base stations, i.e. placing a threshold at the mid point (0.5) between the two base stations. Fig. 2 shows the simulated delay performance (explained in more detail in the sequel) when load split between the base stations is varied from 0.5 (even division of load) to 0.1 (highly asymmetric load division). The results show that the optimal load division depends on the intensity of the offered load, and is not balanced but significantly asymmetric. As exhibited in the figure, where mean delays are plotted on a logarithmic scale, the performance implications can be substantial; load balancing may achieve mean delays 15 times higher versus an optimal asymmetric split. Further, as we will see in the sequel, a user association policy that tries to balance loads can, in certain cases, even result in Braess's paradox like phenomena where the addition of extra resources (base stations) can result in performance degradation. These results are surprising, and reveal the complexity and substantial impact that dynamic coupling can have in the context of wireless networks. This motivates the need for careful analysis that we will carry out in this paper, as well as comparisons with more complex user and system greedy dynamic policies.

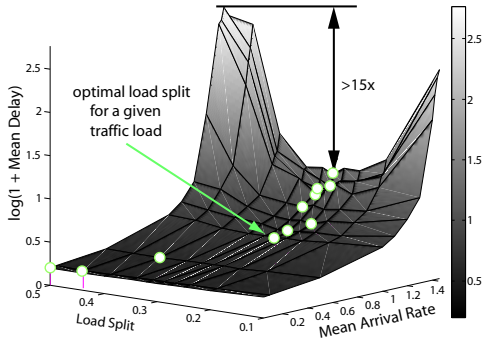


Fig. 2: Mean file transfer delay (on a logarithmic scale) versus different load split thresholds on mobile terminal association and mean traffic loads between the two base stations.

Related Work: Various dynamic policies that split load among base stations have been proposed for different contexts. For example, load balancing schemes have been proposed for the scenario where frequency reuse is used to protect against inter-cell interference, and where the traffic carried by the network is voice [4]–[6]. The objective in these works has been to ensure that load is balanced among base stations.

This philosophy has also been used in addressing the case of best-effort traffic. When the wireless network is subject to spatially heterogeneous traffic loads, emphasis has been placed on the development of schemes that try to balance the load across base stations. Centralized schemes to jointly balance loads and schedule packets are presented in [7]. Such centralized schemes however, incur excessive communication and computational overhead. In [8], a load balancing scheme that requires much less coordination is considered. The scheme tries to explicitly balance the load across base stations, taking into consideration both the long-term rate at which users can be served, and their load. Another idea that was proposed, also in [8], is to lower the strength of the pilot signals that heavily loaded base stations broadcast, so as to discourage users from joining them. A scheme that is similar in spirit is proposed in [9], called MAC-cell breathing, that attempts to balance the load in all base stations. The above-mentioned schemes however, assume implicitly that the base stations in the network are always transmitting and thus interfering with transmissions in their neighboring cells. The focus in these schemes is to ensure that the load being served by different base stations in a neighborhood is as similar as possible. In [10], the cell geometry that results from users joining the base station that offer the highest uplink signal to interference plus noise ratio is studied.

The case of dynamic traffic, with the associated bursty interference, has not been extensively studied. In [11], the effect of equalizing the load in neighboring base stations was studied through simulation, and it was observed that load balancing did not make much of a difference under heavy load. This problem is also studied in [12], but under the assumption that transmissions are orthogonal. The impact of dynamic interference was also demonstrated in [13], wherein the problem of load balancing in a hybrid wireless local area/wide area network was studied using approximations proposed in [14].

The stability region of a dynamic system with interacting servers under load balancing strategies was examined in [15]. The stability region was explicitly characterized in the case of a two-server system, and a lower bound on the stability region was obtained for systems with multiple servers. The stability region in the case of static load balancing policies and a class of dynamic policies was also studied in [16]. A surprising result is that the stability region of the system is not always maximized by perfect load balancing across servers. While the above papers address the question of determining the network capacity, they do not provide insight into designing user association policies to optimize performance perceived by users in a system serving a load that is in the interior of the stability region. In contrast, the focus of this paper is on designing practical user association policies that optimize flow-level performance, i.e., the actual file transfer delay experienced by users.

Our contributions: In addition to unpredictable short-term 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. We present two interference-sensitive user association schemes that attempt to optimize user-perceived file transfer delays: an optimized static scheme obtained through solving a semidefinite optimization problem that uses prior knowledge of long-term spatial loads and a measurement-based policy that infers the nature of interference and spatial load and adapts to it. Our contributions in this context include:

1. We propose a methodology to optimize the performance of wireless systems coupled through dynamic interference and

apply it to study networks with base stations distributed on a line and on a two dimensional plane. To our knowledge, prior to this work, no closed-form results or good approximations were available for general systems with 3 or more base stations serving dynamic traffic loads.

2. For a dynamic model of the user association problem in one dimension, we show that delay optimal static policies are threshold based. Surprisingly, we find that even for a symmetric network, a policy which balances load can be highly suboptimal. Moreover, we find that asymmetric policies can improve average delays seen by users at *all* spatial locations.

3. We show that an optimized static policy (asymmetric) can substantially outperform dynamic policies which are greedy from the user's or system's points of view and achieves performance close to that of a 'repacking' policy. This suggests that an important objective for protocol and network design will be to achieve such asymmetries.

4. We present ISAP (Interference-Sensitive, Adaptive Policy), a novel load association policy based on inducing asymmetry in the load carried by base stations that uses measurements to infer the degree of performance coupling due to inter-cell interference, and adapts to it.

5. We demonstrate through extensive simulations that the proposed policies consistently outperforms conventional, load balancing based approaches under both spatially homogeneous and heterogeneous loads. These results also show that the performance of conventional dynamic schemes is highly dependent on the spatial load, and no single best scheme can be identified. In certain extreme scenarios, we show that these conventional user association policies can even result in Braess's paradox like phenomenon.

Our results exhibit the importance of understanding the impact of dynamic traffic and interference in wireless networks.

Organization of paper: The system model is described in detail in Sec. II. The optimal static association policy is characterized in Sec. III, while Sec. IV explores the impact of asymmetric static association policies. The methodology used to pick the optimal static policy is presented in Sec. V. Simulation results comparing the performance of the static policy to various dynamic strategies is presented in Sec. VI.ISAP, an adaptive policy for load allocation across interfering base stations is presented in Sec. VII. Simulation results characterizing the performance of the various policies under non-homogeneous spatial loads are presented in Sec. VII-D. The sensitivity of delay performance to file size distributions and system and channel parameters is considered in Sec. VIII, while Sec. IX explores a scenario where adding resources (base stations) results in deterioration of delay performance. Sec. X concludes the paper.

II. System Model

In Secs. III-IV, we consider two base stations, BS1 and BS2, located a distance d apart on a line, as shown in Fig. 1a. User requests are distributed on the line segment joining the two base stations. We identify a user request by the distance between the user and BS1, denoted by $x \in [0, d]$. The distance between the user and BS2 is then given by $d - x$. User requests arrive according to a spatial Poisson process with mean measure $\lambda(\cdot)$ which is absolutely continuous with respect to the Lebesgue measure, i.e., the rate at which user requests arrive into a set $\mathcal{X}$ is $\lambda(\mathcal{X})$. We assume that each user request corresponds to a downlink file transfer which is assumed to be exponentially distributed with mean 1, and the position of the user remains fixed for the duration of the transfer. Once the file transfer is completed, the user leaves the system.

The capacity to users from their serving base station depends on the received signal strength and the strength of the received interference, and is assumed to be monotonically increasing in the perceived signal to interference plus noise ratio (SINR). The base stations transmit, and thereby cause interference only when they are serving users. We assume that the base stations use the processor sharing mechanism to serve active users, i.e., the base station splits time evenly among all users currently being served. Thus, a degree of temporal 'fairness' is imposed.

We classify user association policies into static and dynamic policies. *Dynamic* policies use information about the current loads being served at the candidate base stations when deciding the base station to which a new user is assigned. A *static* user association policy is one that does not take into account the current state of the system when making this decision. A static load allocation policy π partitions the line segment into regions $\mathcal{X}_1^\pi$ and $\mathcal{X}_2^\pi$, served by BS1 and BS2 respectively. The base station that serves a user at location x under policy π is denoted by $\beta^\pi(x)$. Thus, if $x \in \mathcal{X}_1^\pi$ then $\beta^\pi(x) = 1$, otherwise $\beta^\pi(x) = 2$. Base stations transmit at maximum power when there are active associated users, and turn off otherwise. The signal strengths received by a user at location x from BS1 and BS2 are denoted by $s_1(x)$ and $s_2(x)$ respectively. For $i = 1, 2$, we denote the worst and best received signals in $A \subset [0, d]$ by $\underline{s_i}(A) = \inf_{x \in A} s_i(x)$ and $\overline{s_i}(A) = \sup_{x \in A} s_i(x)$. Let N_0 denote the average power of the additive Gaussian noise.

Under a given policy π, we let $\mathbf{U}^\pi(t) = (\mathcal{U}_1^\pi(t), \mathcal{U}_2^\pi(t))$ where $\mathcal{U}_i^\pi(t)$ is the set of locations for users being served at base stations $i = 1, 2$ at time t. Note that since $\lambda(\cdot)$ is non-atomic, users' locations will be distinct with probability 1. Given our assumptions on arrivals and file sizes, $\mathbf{U}^\pi(t)$ is a Markov process for static user association policies π since, given all the users locations, one can determine their service capacities and thus departure rates. Note however that its state space is uncountable. By contrast, the process $\mathbf{Q}^\pi(t) = (Q_1^\pi(t), Q_2^\pi(t))$ defined by $Q_i^\pi(t) = |\mathcal{U}_i^\pi(t)|$ for $i = 1, 2$ is on a countable state space, but not Markovian.

This model is similar to that of optimally routing n classes of users to m non-identical queues studied in [17], with an infinite number of classes. However, in our case the problem is further complicated by the fact that the queues at the base stations are coupled (through interference) and the system is non-work-conserving. Systems of coupled queues have been analyzed in the past [18]–[21], but the problem is extremely difficult. Closed form expressions are known only in the case of some simple work-conserving scenarios with two coupled queues and only asymptotic results are known for more general cases. [18]–[21] Even the problem of characterizing the stability of coupled queues which was addressed in [22]

is difficult, and one has to employ numerical methods.

A. Simulation Model

We simulate a one-dimensional network consisting of two base stations located 500m apart as well as a network on a plane consisting of three facing sectors in a hexagonal layout of base stations with cell radius 250m. User requests arrive according to a Poisson process. In Sec.VI, we simulate a user distribution that is spatially homogeneous. In the two base station case, users are assumed to be distributed uniformly on the line joining the two base stations, and in the three base station network, users are assumed to be distributed uniformly within the hexagon formed by the three interfering sectors. We consider non-homogeneous spatial load distributions in the simulation results presented in Sec. VII-D. and the exact load profiles simulated are described therein.

A carrier frequency of 1GHz, and a bandwidth of 10MHz are assumed. The maximum transmit power is restricted to 10W. Additive white Gaussian noise with power -55dBm is assumed. We consider a log distance path loss model [23], with path loss exponent 2. File sizes are assumed to be exponentially distributed, with mean 5MB. The data rate at which users are served is calculated based on the perceived SINR using Shannon's capacity formula. The maximum rate at which a user can be served is capped at 54 Mbps. The base stations transmit at maximum power when they have active users, share capacity across users using a processor sharing mechanism, and turn off otherwise. The mean user-perceived delay is estimated within a relative error of 2%, at a confidence level of 95%. Note that the sensitivity of the delay performance to the channel and system model is examined in Sec. VIII where a system with a higher path loss exponent, and cell-edge SNR of 10 dB is simulated.

III. Optimal Static Policies

We begin by considering static association policies in the one-dimensional, two base station system. Such policies are defined by the service regions corresponding to each base station, which in turn may depend on the long-term offered load $\lambda(\cdot)$. The key result is that under our system model, the service regions are contiguous and thus are defined by a single threshold between the two base stations. The following lemma provides a partial characterization of optimal static policies. Note at the outset that, while this result appears straightforward, the challenge lies in the dynamic nature of the model; specifically, in dealing with the spatial arrivals and departures, the dynamic (on/off) nature of the interference from the neighboring base station, and thus the coupling of delay performance between the two base stations.

Lemma 3.1: Consider the two base station model defined in Sec. II. For any static load allocation policy π_a with $\mathcal{R}_1 \subseteq \mathcal{X}_1^{\pi_a}$, $\mathcal{R}_2 \subseteq \mathcal{X}_2^{\pi_a}$ with $\lambda(\mathcal{R}_1) = \lambda(\mathcal{R}_2)$, and such that $\underline{s_1}(\mathcal{R}_2) \geq \overline{s_1}(\mathcal{R}_1)$ and $\overline{s_2}(\mathcal{R}_2) \leq \underline{s_2}(\mathcal{R}_1)$, the policy π_b with $\mathcal{X}_1^{\pi_b} = (\mathcal{X}_1^{\pi_a} \cup \mathcal{R}_2) \setminus \mathcal{R}_1$, $\mathcal{X}_2^{\pi_b} = (\mathcal{X}_2^{\pi_a} \cup \mathcal{R}_1) \setminus \mathcal{R}_2$ achieves lower (or equal) average user delay.

The insight underlying this lemma can be grasped by considering Fig. 3. It illustrates a policy π_a which satisfies the lemma's conditions if signal strength decays monotonically with distance from the serving base station – although part of our system model, this is not required to prove the lemma. Policy π_b is constructed by merely exchanging service regions $\mathcal{R}_1, \mathcal{R}_2$ between the two base stations. The constraints on the best and worst case signal strengths ensure that this exchange is favorable for both base stations at all the associated user locations, which implies the following straightforward fact.

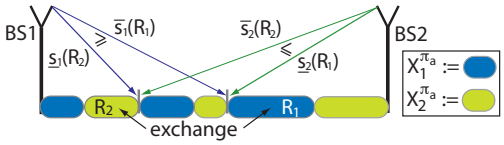


Fig. 3: A sub-optimal load allocation policy.

Fact 3.1: Under the assumptions on $\mathcal{R}_1$ and $\mathcal{R}_2$ in Lemma 3.1, and the assumption that capacity is monotonically increasing in SINR, the capacity from BS1 to any user in $\mathcal{R}_2$ is greater than that to any user in $\mathcal{R}_1$ under the *same* interference regime, i.e., BS2 is transmitting or not. Similarly, the capacity from BS2 to any user in $\mathcal{R}_1$ is greater than that to any user in $\mathcal{R}_2$, whether BS1 is transmitting or not.

So, the exchange leaves the intensity of arrivals to BS1 and BS2 unchanged, and associates users to them which then can be served at higher capacity under the same interference regime. This allows us to construct a spatial coupling (i.e., by associating users in different regions) for networks under the two policies, showing that the average queue lengths are not increased. The details of this argument are in the appendix, and can be extended to other service disciplines, e.g., FCFS and LCFS.

Theorem 3.1: For the two base station model defined in Sec. II, there exists a static load allocation policy minimizing mean delay corresponding to a spatial threshold $x^* \in [0, d]$ such that a user at location x is served by BS1 if $x \leq x^*$ and by BS2 otherwise. This can also be expressed as a threshold on the ratio of received signal strengths from the two base stations.

Proof Sketch: Since traffic intensity measure $\lambda(\cdot)$ is non-atomic, if the service regions associated with the BS1 and BS2 are not contiguous, one can construct regions $\mathcal{R}_1$ and $\mathcal{R}_2$ satisfying Lemma 3.1. Thus, a new policy can be constructed by exchanging regions $\mathcal{R}_1$ and $\mathcal{R}_2$ between the base stations' service regions without increasing the mean delay. This exchange operation can be repeated as long as the service areas are not contiguous. Thus an optimal policy must be defined by contiguous regions, i.e., specified by a spatial threshold. Since the ratio of the received signal strengths is strictly decreasing or increasing with the received signal strength (or distance) from a base station, the policy can also be implemented as a threshold on this ratio.

Note that optimal static load allocation policies need not necessarily be unique. For example, consider the case when user requests are distributed homogeneously on the line segment joining the two base stations. If the optimal threshold does not correspond to the midpoint, then by symmetry, the policies that divide the service areas using thresholds at a

distance d^* from BS1 and d^* from BS2 will result in identical mean user delays.

IV. Optimal Threshold Trends

As a consequence of Theorem 3.1, we need only consider threshold-based static allocation policies. Fig. 2 exhibits the simulated mean user delay for varying thresholding policies as the (spatially homogeneous) arrival rate between the base stations increases. The policies are characterized by the fraction of load served by BS1 with 0.5 corresponding to load balancing and 0.1 to only 10% of the load. Due to symmetry, the delay performance would be identical if the threshold were moved closer to BS2. For each arrival rate, the optimal load split, i.e., roughly achieving the minimum mean delay, is highlighted. We make the following observations:
1. The location of the optimal threshold is a function of the load on the system.
2. Except at very low loads, delay performance is improved by moving the threshold away from the mid-point, thus inducing asymmetrical loads on the two base stations.

Why does this happen? Load balancing increases parallelism, i.e., base stations are more likely to be simultaneously active. In our model, load balancing associates users with close by base stations providing them a stronger signal. Finally, it would appear that load balancing might be beneficial in terms of statistical multiplexing at the two base stations. If capacity users see was fixed, these points would provide the right insight. Yet, when dynamic interference is present, the capacity users see (particularly those far from either base station) can be substantially reduced by interference, and the fraction of time that base stations interfere with each other depends on the traffic and the load allocation policy. Thus, when arrival rate is low, the probability of the base stations being simultaneously active is low; the base stations operate in an interference-free environment, and load balancing is roughly optimal. For higher arrival rates, performance is strongly impacted by interference, and skewing the load is beneficial. Intuitively, this skew reduces the utilization of one of the base stations, say BS1, and thus the interference it causes on BS2's users, which reduces BS2's utilization, in turn benefiting BS1. However, one cannot overdo this skew as serving users that are far away, and thus have poor received signal, is also detrimental. Finally, it is tempting to assume that as load increases, base stations are always busy and the role of dynamic coupling reduces. Yet, as can be seen, at high loads performance sensitivity is also high, and the gains of an optimal asymmetric split increase further. The optimal threshold reflects a complex tradeoff among dynamic interference, statistical multiplexing, and users' signal strengths.

V. Optimizing the Threshold

In this section, we propose an approximation methodology for optimizing static load allocation policies for the wireless network model in Sec. II, naturally extended to N base stations serving a possibly higher dimensional region. A policy π partitions the service area such that base station n has service area $\mathcal{X}_n^\pi$ and overall arrival rate $\lambda_n = \lambda(\mathcal{X}_n^\pi)$. First, we approximate the Markovian model with uncountable state space by one with a countable state space, i.e., we will no longer keep track of the locations of users associated with each base station. This involves introducing an 'effective' rate for *all* users associated with a base station which depends on the busy state of the remaining base stations. Thus, the model preserves the dynamic interference characteristics. Then, we use a semidefinite programming based approach to upper/lower bound the performance for this approximated model. Finally, we propose optimizing performance over families of static policies that can be easily parametrized, e.g., for our one-dimensional example, one need only determine the threshold. The subsequent section shows that the accuracy of the proposed methodology is excellent.

A. Countable State-Space Approximation

We let $\vec{Q}(t) = (Q_n(t), n = 1, \ldots, N)$ denote the number of active users at each base station at time t for our approximated process that models the evolution of the number of users being served by each base station. For notational simplicity, we have suppressed its dependency on π. As mentioned earlier, the capacity to a user depends on *both* its current location and the interference profile it sees from neighboring base stations. We let $\vec{\Delta}(t) = (\Delta_n(t), n = 1, \ldots, N)$ where $\Delta_n(t) = \mathbf{1}(Q_n(t) > 0)$ denotes the status (idle or busy) or the 'interference profile' of the base stations. Note that $\vec{\Delta}(t)$ can take 2^N possible values which we denote $\vec{\delta}^i, i = 1, \ldots, 2^N$. Let $c_n(x, \vec{\delta}^i)$ denote the actual capacity at which base station n can serve a user at location $x \in \mathcal{X}_n^\pi$ under interference profile $\vec{\delta}^i$.

A user's sojourn time, i.e., The time that users spend in the system is inversely proportional to their service capacity. Thus, the mean rate at which users in a cell can be served depends on the steady-state distribution of users that is induced in the cell (which differs from the distribution of arrivals). As shown in [24], the effective service capacity of a base station is given by the harmonic mean of the user capacities, when these are not time-varying. In our approximate model, the effective capacity under interference profile $\vec{\delta}^i$ depends *only* on $\vec{\delta}^i$ and is given by

$$c_n^{\vec{\delta}^i} = \left(\int_{\mathcal{X}_n^\pi} \frac{1}{c_n(x, \vec{\delta}^i)} \frac{\lambda(dx)}{\lambda_n} \right)^{-1},$$

the *harmonic mean* of the users service capacities under $\vec{\delta}^i$ weighted by the spatial distribution of arrivals to the base station, i.e., $\frac{\lambda(dx)}{\lambda_n}$. Since, in reality, each user does observe different rates over the course of time depending on the activity level of the neighboring base station, these effective capacities are an approximation. However, users with low received signal strength tend to be located near the cell edge, and are also typically subject to high levels of inter-cell interference. As a result, the users at locations that receive comparatively low service rates in the interference profiles without interference also receive lower service rates under the other interference profiles. Thus, in most cases, we expect this approximation to be reasonable. Since files have mean size of 1, the total service rate $\mu_n^{\vec{\delta}^i}$ at base station n under interference profile $\vec{\delta}^i$ is given by $\mu_n^{\vec{\delta}^i} = c_n^{\vec{\delta}^i}$. We assume that the system is stable

and let μ^* denote an upper bound for the maximum service rate for any base station.

Our approximation is given by a continuous-time Markov process with transition rate bounded by $\eta = \sum_{n=1}^{N} \lambda_n + N\mu^*$, so it can be uniformized. With a slight abuse of notation, we let $\vec{Q}(k)$ denote the state for the uniformized discrete-time Markov chain and $\vec{\Delta}(k)$ the associated interference profile at discrete-time step k. The transition probabilities for the uniformized Markov chain are as follows. Suppose $\vec{Q}(k) = \vec{q}$ has associated interference profile $\vec{\delta}^i$, i.e., $\delta_n^i = \mathbf{1}(q_n > 0)$ then

$$\mathbf{P}(\text{arrival to queue } n | \vec{Q}(k) = \vec{q}) = \frac{\lambda_n}{\eta},$$

$$\mathbf{P}(\text{departure from queue } n | \vec{Q}(k) = \vec{q}) = \frac{\mu_n^{\vec{\delta}_i}}{\eta} \delta_n^i,$$

$$\mathbf{P}(\text{no change} | \vec{Q}(k) = \vec{q}) = 1 - \frac{\sum_{n=1}^{N} \lambda_n + \mu_n^{\vec{\delta}_i} \delta_n^i}{\eta}.$$

Note that, if it exists, the uniformized chain's stationary distribution is identical to that of the original. Also, its evolution can be represented as a stochastic recursion

$$\vec{Q}(k+1) = \vec{Q}(k) + \vec{X}(k), \quad k = 0, 1, \ldots,$$

where $\vec{X}(k) = (X_n(k), n = 1, 2, \ldots, N)$ denotes increments in the queues. An arrival into queue n at iteration k is represented by $X_n(k) = 1$, a departure by $X_n(k) = -1$ and if the transition corresponds to the self-loop, $\vec{X}(k) = \vec{0}$. Note that $\vec{X}(k)$ and $\vec{Q}(k)$ are not independent, e.g., one can not have a departure from an empty queue. However they are clearly conditionally independent given $\vec{\Delta}(k)$. For systems with this property, bounds on the mean sum queue length (and other metrics) can be obtained using a semidefinite programming approach, see [25], [26]. Bounds on the mean queue lengths in turn translate to bounds on the mean delay via Little's Law.

B. Determining Optimal Thresholds

As mentioned earlier, when policies can be easily parameterized, one can use these bounds to optimize performance. For our two base-station scenario, Theorem 3.1 shows the optimal static load allocation policy is determined by a simple threshold. The threshold determines the arrival rate of user requests to each base station, which can be obtained simply by integrating over each base station's service region. Also, the effective service rate at each base station can be determined using the approximations described in the previous section. Thus, the transition probabilities for the Markov chain of base station queue lengths can be determined. So for any threshold, the semidefinite programming approach developed in [25], [26] can be used to determine bounds on the mean delay, and a simple line search can be used to determine the threshold giving the smallest lower bound on the mean delay. In the case of the three base station network considered in the sequel, we parametrize policies based on weights associated with the base stations, as described in Sec. VI-C.

Fig. 4 exhibits the computed approximate optimal thresholds versus those obtained via brute-force simulation for our two base station model. As can be seen, both load splits (thresholds) and resulting mean delay performance are very close, supporting the accuracy of our optimization methodology. The optimization approach also provides the flexibility to address complex traffic loads as well as systems with a larger number of base stations as we will see in the sequel.

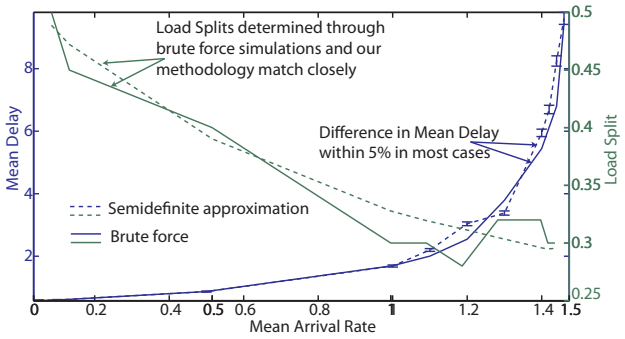


Fig. 4: Goodness of optimized thresholds.

VI. Performance Comparison

A. Comparing Static Policies

Fig. 5-i again illustrates the impact that the choice of threshold location has on delay performance. The user distribution is spatially homogeneous, so locating the threshold at the midpoint between the base stations corresponds to a static load balancing approach. As can be seen, the resultant mean user delays are greatly decreased by choosing an optimal threshold, particularly at moderate to high system loads. These results match the ones obtained in [11] where the authors found that load balancing only increased capacity in the case of large cells where interference did not have a significant impact, and was inefficient in interference-dominated scenarios. The intuitive explanation for this, as presented in [11] as well, is that load balancing tends to increase the simultaneous utilization of base stations thus increasing interference. Also, in [16], load balancing schemes that maximize the stability region of interacting servers was considered. It was found that the load balancing did not maximize the stability region. Here, we can see that a similar result holds from the point of view of performance and the extent of asymmetry required to optimize performance has been quantified. Fig. 5-ii further exhibits the spatial distribution of user delays under the two schemes when the rate at which user requests arrive in the network is 1.2 per second. Surprisingly, skewing the load towards one base station does not result in a trade off where a subset of the users, e.g., at the heavily loaded base station, experience poor performance. Instead, under the optimal policy, the overall impact of inter-cell interference is reduced such that all users, irrespective of their spatial location or perceived signal strength, see improved performance on average.

B. Optimized Policy vs. Dynamic Strategies

Next we compare the performance of the optimal static policy versus the following three dynamic policies:

Greedy User: each new user joins the base station which

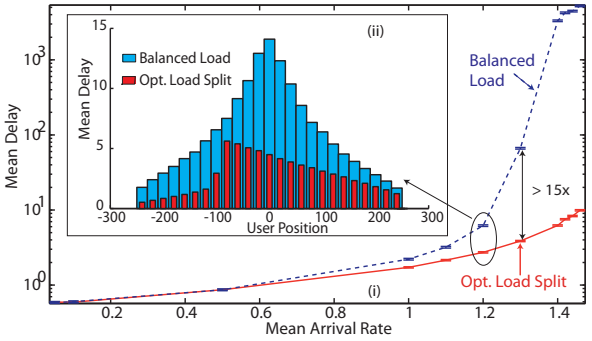


Fig. 5: Mean delay performance (i) and spatial delay distribution (ii) of the optimized policy vs. static load balancing.

offers the highest current service rate. This requires knowledge of the new user's capacity to each base station when the neighbor is active/idle and the number of users each is serving.

Greedy System: each new user is assigned to the base station so as to maximize the resulting current sum service rate of the base stations. This policy is more complex than the Greedy User policy as in addition it requires knowledge of the capacity for *all* ongoing users with and without interference.

Repacking: each time a user arrives or leaves, the assignments of *all* users are chosen so as to maximize sum service rate of the base stations via a brute-force search – the overheads and complexity of such a scheme would be unrealistically high, yet we hypothesize that it results in the best delay performance among non-anticipative dynamic schemes, since at any given time it minimizes the expected time to the next departure.

Fig. 6a illustrates the mean delay (logarithmic scale) for varying traffic loads under the above-mentioned greedy policies. Surprisingly, the optimal static policy substantially outperforms the two greedy polices at moderate to high loads. Indeed, at high load, the mean delay of the static policy is 6 times lower than the greedy system policy which itself is orders of magnitude lower than the greedy user policy. As expected, the repacking policy shown in Fig. 6b (linear scale) is the best, but indeed very close to the optimal static policy.

Fig. 6c exhibits the spatial delay distribution under the system-level greedy scheme vs. the static policy. While the greedy policy exhibits perhaps desirable spatially symmetric performance, it is still the case that the optimal static policy gives better performance to all user locations.

C. Three Base Station Network

The three base station case can be used as a building block to develop a load allocation policy in a larger network. The number of base stations that can potentially serve a particular user request is unlikely to be very large. A load association policy that decides only between the three strongest base stations for each user request seems to be a reasonable tradeoff between complexity and performance. For the 2 dimensional three base station network described in Sec. II-A, the form of the optimal static association policy is difficult to characterize. We compute the 'optimal' static association policy within a family of policies that can be easily parametrized.

1) Weighted signal strengths: The first family of policies we consider is parametrized by base station weights. Each base station is assigned a weight and a user is associated with the base station that offers the maximum weighted received signal strength. The weight associated with one of the base stations is set to 1, and a simple gradient descent is used to determine weights for the remaining base stations. The semidefinite based programming based bounding methodology described in [25], [26] is used to approximate the mean delay at each step of the gradient descent algorithm.

If the base stations are part of a larger network, accurately accounting for the activity levels of other neighboring base stations could be important. The proposed methodology can still be used in such a scenario by including queues corresponding to the neighboring base stations when the performance bounds are computed using the semidefinite optimization. The objective function would remain the expected sum queue length at the three sectors under considerations. Note that including additional base stations will increase the complexity of the bounding procedure.

2) Pairwise optimization: As an alternative to the methodology proposed above, we consider a family of policies where modifying a single parameter while keeping the rest constant allows the load division between two base stations to be modified without affecting the set of users served by the other base station. Note that the policy presented in Sec. VI-C1 does not possess this property as changing the weight associated with any base station potentially changes the load served by all three base stations. This property allows the sequential optimization of the policy parameters, and the optimal policy can be determined using a sequence of iterations where one parameter is adjusted in each iteration. This is particularly important if additional neighboring base stations have to be taken into account. When one of the parameters is being optimized, the only sectors that have to be considered in the optimization are the neighbors of the two base stations that are affected. Thus, each parameter can be optimized while accounting for a different set of neighbors. This reduces the complexity of the semidefinite program that has to be solved to obtain the performance bounds.

The vector of received signal strengths from the three base stations, $\vec{s}(x) = (s_1(x), s_2(x), s_3(x))$, is projected down on to the two dimensional hyperplane that passes through the origin and is orthogonal to the vector $(1,1,1)$. The family of static policies that we consider divide this hyperplane into regions, and a base station serves all users whose projected signal strength vector falls in its region. The hyperplane is chosen such that users with identical relative received signal strengths from the base stations are mapped to the same point. The projected vector, after an orthogonal transformation is given by $\vec{z} = \{z_1, z_2\}$, where

$$z_1 = \frac{1}{\sqrt{6}}(2s_1(x) - s_2(x) - s_3(x)) \text{ and } z_2 = \frac{1}{\sqrt{2}}(s_2(x) - s_3(x)).$$

The hyperplane is divided into three regions by three rays extending from the origin, as shown in Fig. 7. Each base station serves the region between two rays as illustrated in the figure. The rays are specified by the angles α, β, and γ that

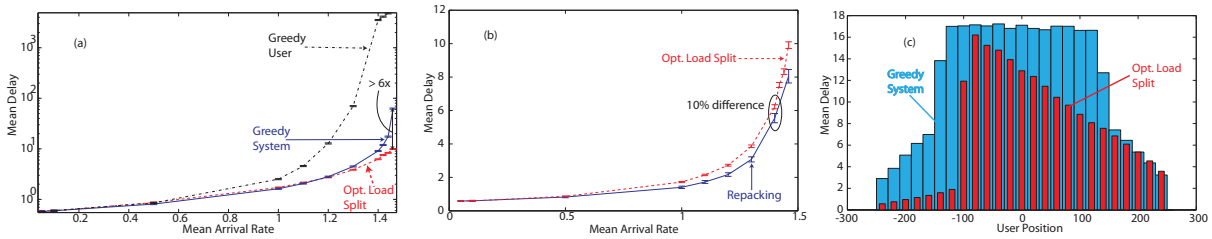


Fig. 6: Comparisons of mean delay for optimal static policy: (a) versus greedy schemes (log scale); (b) versus repacking scheme (linear scale); (c) versus greedy system in terms of spatial delay distribution.

they subtend with the z_1 axis, and these angles parametrize a policy within the family. Rotating one of the rays only exchanges load between the two base stations whose service regions adjoin the ray. The optimal static policy is determined through a series of iterations. At each iteration, one of the parameters is modified, and a new value that improves the overall delay experienced by the set of users served by the three base stations is chosen. Thus, each iteration lowers the overall mean delay experienced by users in the system, ensuring that the optimization procedure converges.

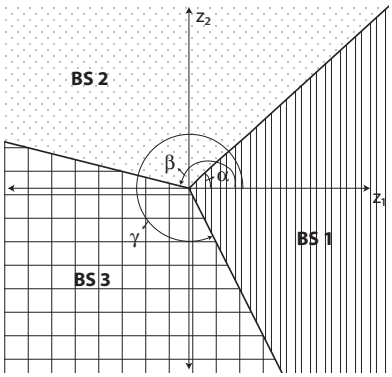


Fig. 7: Load division after projecting down to the two dimensional hyperplane

Fig. 8 exhibits the mean delay performance in a three base station network. The repacking policy for this case is a hard combinatorial problem to be solved upon each arrival/departure and so was infeasible. The static load balancing and the greedy user policies exhibit similar performance, i.e., overlap, while the optimized static (asymmetric) policy exhibits substantial performance gains. Even the greedy system policy (itself unrealistic in practice) achieves mean delays up to 20 times higher than the weighted signal strength based policy. The projection based policy performs significantly better than even the weighted signal strength policy, reducing the user-perceived mean delay further by 6-10 times at high loads. The projection based policy is sensitive only to variations in the relative received signal strengths that users perceive from the three base stations, and the metric used to associate users with base stations is no longer linearly dependent on the users' signal strengths. The results demonstrate that such a policy when optimized, taking into account the effect of coupling in the wireless system is indeed effective at achieving good delay performance. The results also suggest that further gains could be attained by a family of policies that allows more flexibility in dividing load across base stations.

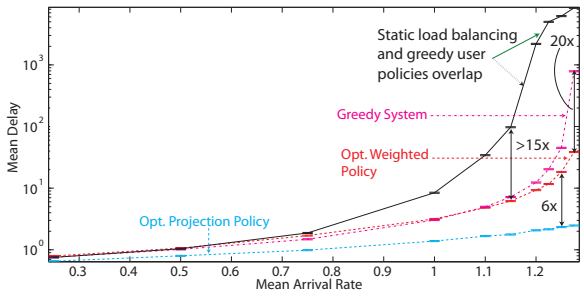


Fig. 8: Three Base Station Network: Mean Delay Performance under the weighted signal strength policy

VII. The Interference-sensitive, Adaptive Policy

The static policy developed in Sec. V requires knowledge of the long-term traffic loads served by the wireless network. Also, several iterations of a semidefinite optimization problem have to be solved in order to determine the optimal thresholding policy. Further, the static policy determined through the optimization procedure may not be robust to quickly changing traffic loads. In this section, we present the Interference-Sensitive, Adaptive Policy (ISAP) that divides load among base stations and induces asymmetry by tracking the impact of performance coupling among base stations resulting from dynamic inter-cell interference. The proposed policy requires no communication among base stations and only requires each base station to track two simple measures of the load being served. We compare the performance of the proposed adaptive policy with the dynamic policies discussed earlier, as well as the optimized static policy which is a useful benchmark.

The load on the system depends not only on the rate at which users arrive and the mean file size requirements but also on their location with respect to the base stations. A load allocation policy must be sensitive to both the intensity of the load as well as its distribution in space. The policy must be able to distinguish between scenarios where inter-cell interference is responsible for causing high user delays, and scenarios where user delays are driven by high traffic loads inherent in the system. As seen in the previous sections, in an interference-dominated scenario, an adaptive scheme may need to create an asymmetric division of load among base stations. Schemes where the desirability of a base station depends solely on the nature of the load supported by that base station will not possess this property.

A. Measuring the impact of interference

In order to estimate the effect of inter-cell interference, each base station i tracks and maintains estimates for the *system load* and *virtual load* as described below. We let $\hat{\rho}_i^S$ denote an estimate of the true *system load*, the current utilization of the downlink queue of base station i. This will reflect the effect of interference from neighboring base station transmissions. We let $\hat{\rho}_i^V$ denote an estimate of the *virtual load*, the base station utilization that would result if base station i experienced no interference. Measurements are updated in discrete-time slots of length δ. In any slot, the base station is either idle or transmitting to exactly one user, say user j. Slots are indexed by $t \in \mathbb{Z}$, corresponding to times δt. The transmission rate to user j in slot t under the current policy, taking into account the current activity state of the neighboring base station(s) is denoted $r_{ij}(t)$, and the rate to the user in the absence of *any* interference is denoted $r_{ij}^0(t)$. Each base station estimates the current system load and the virtual load resulting from the current load allocation policy as follows.

Estimating the system load: The system load can be estimated by periodically checking if there are active users associated with the base station. Each base station updates the estimate for the system load at δ intervals as follows:

$$\hat{\rho}_i^S(t+1) = \beta_S \mathbf{1}(\text{BS } i \text{ is transmitting in slot } t) + (1-\beta_S)\hat{\rho}_i^S(t),$$

where $\beta_S \in (0,1)$ is a small constant determining the averaging time scale. Note that we assume time slots are small enough that base stations are either on or off for the entire duration of a slot.

Estimating the virtual load: A base station's virtual load is measured as the fraction of time the base station would be actively transmitting to users if it were to serve the same traffic in the absence of interference. The estimate for the virtual load is updated along with the system load as follows:

$$\hat{\rho}_i^V(t+1) = \beta_V q_i^V(t) + (1-\beta_V)\hat{\rho}_i^V(t),$$

where $\beta_V \in (0,1)$ is a small constant, and the function $q_i^V(t)$ is defined as

$$q_i^V(t) = \begin{cases} 0, & \text{BS } i \text{ is idle} \\ \frac{r_{ij}(t)}{r_{ij}^0(t)}, & \text{BS } i \text{ is transmitting to user } j \end{cases}$$

One can interpret the virtual load as follows: The virtual system serves exactly the same user as the real system in each slot. The virtual system transmits exactly the same number of bits to the user as the real system, by using only a fraction $\frac{r_{ij}(t)}{r_{ij}^0(t)}$ of the slot. Thus, when the user in the real system experiences interference in a slot, the slot is only partially used in the virtual system and the base station is idle for the remainder of the slot.

Thus, the virtual system is not work conserving. However, in the case where the channel to the users is time-invariant, the fraction of time that the base station is busy transmitting in the virtual system is equal to the utilization of the base station under any work conserving policy in a hypothetical system with no interference . To see this, note that one can rearrange the times at which the users are served in the virtual system to match any work conserving policy. In the case of time-varying user rates, as long as the rate $r_{ij}^0(t)$ is stationary and the scheduling of data in the real system is not channel-aware, e.g., processor sharing, and thus independent of the channel capacity under no interference, the fraction of time the base station is busy in the virtual system corresponds to the case of the base station in a hypothetical system with no interference. Note that this is not true in general, in the case of time-varying channels and arbitrary scheduling disciplines. However, a similar virtual system could hypothetically be constructed for such cases also.

Estimating the impact of interference: Clearly, $\hat{\rho}_i^S$ will always be greater than $\hat{\rho}_i^V$. The overall impact that inter-cell interference has on base station i can be measured by a function of both $\hat{\rho}_i^S$ and $\hat{\rho}_i^V$, such as $(\hat{\rho}_i^S - \hat{\rho}_i^V)$ or $\frac{\hat{\rho}_i^S}{\hat{\rho}_i^V}$.

B. Algorithm to determine the serving BS

1) Two base station scenario: We begin by considering the two base station case. Suppose a request from user j arrives at slot t, and is to be assigned to one of base stations $i = 1, 2$. Our proposed ISAP policy is a simple weighted maximum rate policy, that is, the user connects to the base station $i^* = \arg\max_{i=1,2} w_i(t) r_{ij}^0(t)$ with ties broken arbitrarily. The novelty in the policy lies in specification of time-dependent weights $w_i(t)$ which in turn are nonlinear functions of the current estimates for the true system and virtual traffic loads seen at *both* of the base stations.

Specifically, we assume the base stations share their current estimates $\hat{\rho}_i^S, \hat{\rho}_i^V$, $i = 1, 2$ with the user. The user in turn assigns a weight of 1 to the base station which currently has the lowest system load. The other base station is assigned a weight which exceeds 1 and increases with the degree of inter-cell interference experienced by the base stations. There are many possibilities for doing this, yet in this paper we consider the following specific weight assignment:

$$w_1(t) = \begin{cases} 1, & \hat{\rho}_1^S(t) \le \hat{\rho}_2^S(t) \\ 10^{(\gamma \prod_{i=1}^2 (\hat{\rho}_i^S - \hat{\rho}_i^V))}, & \hat{\rho}_1^S(t) > \hat{\rho}_2^S(t) \end{cases}$$

where γ is a parameter that can be tuned. The weight assigned to base station 2 is similarly computed. The idea underlying this weight assignment is as follows. If $\hat{\rho}_i^S \approx \hat{\rho}_i^V$ for *one or both* of the base stations then the weight's exponent is roughly 0 and the base station with heavier load will have a weight which is only slightly larger than 1. In this case the policy reduces to a greedy max rate policy. However, if *both* base stations are subject to interference, then $\hat{\rho}_i^S - \hat{\rho}_i^V > 0$ for $i = 1, 2$. So the weight associated with the heavier loaded base station is quickly increasing with the degree of interference seen by base stations's users, and larger than 1. In this case, the policy becomes a weighted max rate policy with a bias towards attracting additional load to the heavier loaded of the base stations, the type of load asymmetry we found to be advantageous earlier in this paper. Clearly other suitable functions of the system and virtual loads are possible, yet the above appears to be reasonable and work well.

2) Multi base station scenario: The proposed policy readily extends to the case of a network with multiple base stations. Suppose a user j arrives at time t and can be associated with any one of n base stations. For simplicity assume the possible base stations are indexed $i = 1, \ldots n$ such that they have decreasing system loads. The multiple base station association policy exhibited in Algorithm 1 relies on making pairwise comparisons among base stations starting from the base stations which are seeing the heaviest loads. The idea is once again to favor asymmetries in load towards base stations which are seeing high system loads, but only if they are also strongly coupled through interference with one of the other base stations.

Algorithm 1 Assigning user j to one of n BSs

Sort the base stations in decreasing order of $\hat{\rho^S}_i(t)$.
Let i^* = 1
for $i = 2$ to n **do**
  $w_{i^*} = 10^{(\gamma(\hat{\rho}^S_{i^*} - \hat{\rho}^V_{i^*})(\hat{\rho}^S_i - \hat{\rho}^V_i))}$
  $w_i = 1$
  if $w_{i^*}(t) r^0_{i^* j}(t) < w_i(t) r^0_{ij}(t)$ **then**
    $i^* = i$
  end if
end for
return i^*

By assigning a weight larger than 1 to the heavily loaded base station, ISAP induces asymmetry in the loads served. However, the asymmetry is controlled as the policy is also sensitive to the rate at which the users can be served by the base stations. Note that if either base station is not affected by interference, the weight associated to the heavily loaded base station will be very close to 1, resulting in a policy that resembles a greedy maximum rate policy.

C. Complexity of ISAP

The individual base stations keep track of the utilization and virtual utilization through one update every slot. When a new user has to be admitted into the system, the user receives the above measures from all the candidate base stations, and uses Algorithm 1 to pick the serving base station. With n base stations, the complexity of the algorithm is equivalent to sorting n values and is $O(n \log n)$.

D. Performance Evaluation

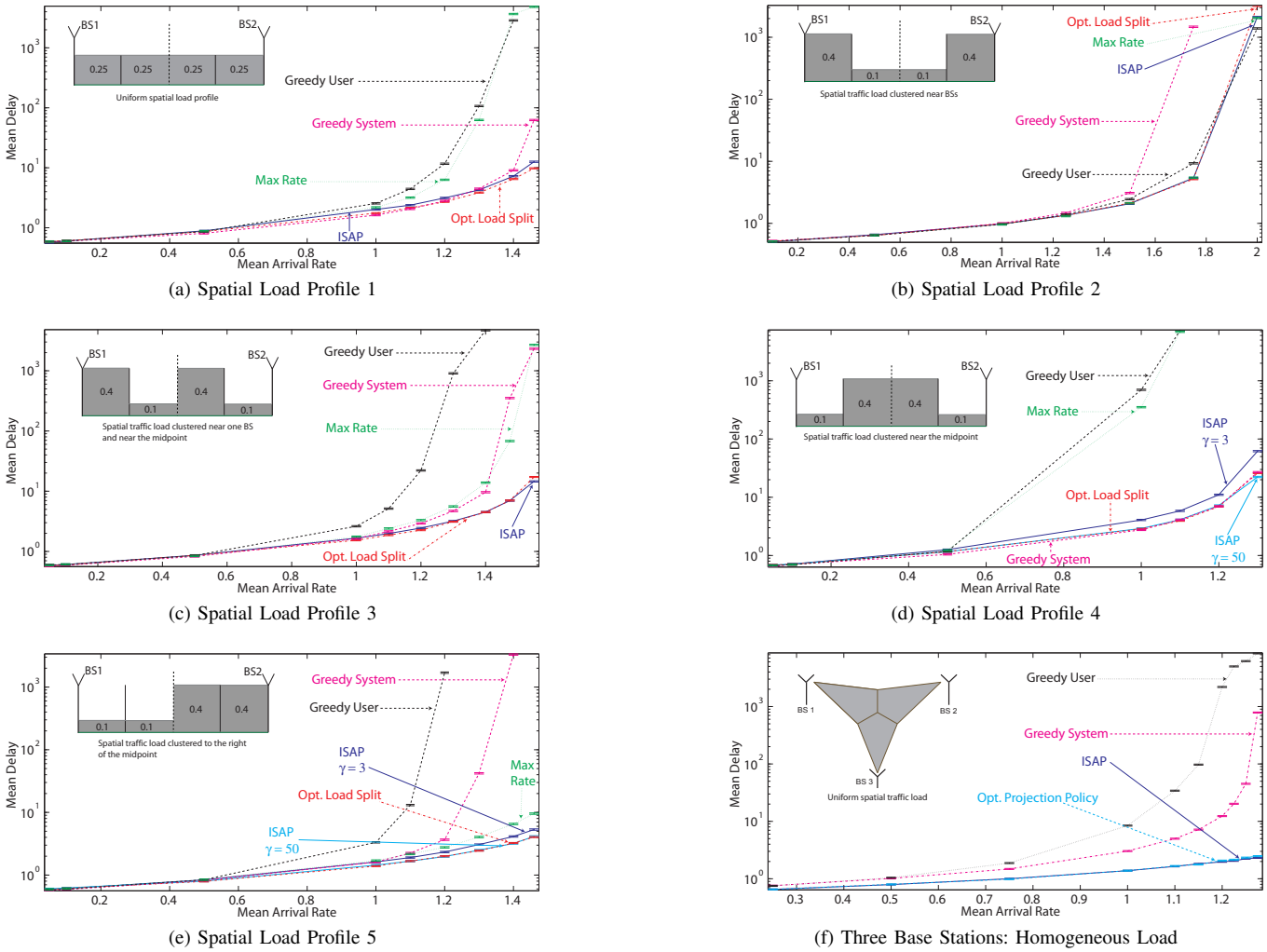
We compare the delay performance of ISAP to the other dynamic schemes introduced in Sec. VI, and the static policy resulting from our approximate SDP based optimization. The value for the constant multiplicative factor γ is 3, unless noted otherwise. This constant should be chosen to ensure that the dynamic range of the weights is sufficiently large.

1) Two base station scenario: Thus far, all the simulation results exhibited performance under spatially homogeneous user (load) distributions. In this section, we additionally consider various spatially non-homogeneous load profiles as shown in Fig. 9a-9e. The line segment joining the two base stations is split into four quarters, and the load distribution is varied by varying the proportion of users in each quarter. Users in a particular quarter are uniformly distributed within that quarter. Load profiles 1, 2 and 4 are symmetric with respect to the midpoint between the base stations. The users are concentrated near the base stations in profile 2, and the impact of inter-cell interference is diminished. The effective load on the network under this profile is lighter at a fixed user arrival rate compared to profile 4, where users are concentrated close to the midpoint and are strongly impacted by inter-cell interference. The load distribution under profiles 3 and 5 is asymmetric.

The optimized static policy and ISAP perform consistently well under all spatial load profiles, and perform as well as or outperform all the dynamic policies. This demonstrates their robustness to spatially heterogeneous traffic loads. Under load profile 3, for example, they outperform all the other schemes by a wide margin. None of the other schemes perform well under all profiles. The proposed schemes are able to infer the nature of the spatial load and adapt to it. Under load profiles 4 and 5, choosing a higher value for the multiplicative constant γ is necessary for the performance of ISAP to match the optimized static policy. Thus, in order to achieve the optimal performance, ISAP has to be parametrized depending on the load distribution. Understanding how to optimize γ is a topic for future study. However, note that even if a nominal value is chosen for γ, ISAP performs very well, even if it is not optimal.

The relative performance of the dynamic schemes can vary dramatically with the distribution of the spatial load. The greedy system scheme performs well under profiles 1 and 4. However, it is the worst among the schemes under load profile 2. Since the greedy system scheme tries to maximize the average throughput realized by all users in the system, it might deviate from a load balancing policy so as to ensure that a base station stays idle. However, since users cannot be reassigned, such decisions adversely affect long-term delay performance. The static max rate scheme performs well under load profile 2, where the effect of interference is minimal and under profile 4, where the spatial load is inherently asymmetric. It performs very poorly under the other spatial profiles.

2) Three (or More) Base Station Network: : Fig. 9f exhibits the mean delay performance that users perceive in a three base station network under ISAP with γ set to the nominal value of 3, the projection based static policy, and the greedy dynamic policies. We only consider the spatially homogeneous user distribution in this case. Evaluating the performance of the various static and dynamic policies under spatially heterogeneous user distributions is a topic for future study. ISAP results in delay performance that is comparable to the projection based static policy. Both policies perform significantly better than the dynamic policies. An alternative way to compare the performance of the policies is through the extra traffic that can be supported while maintaining mean user-perceived delay under a particular threshold. For example, considering a threshold of 1 sec. for the user to be able to download the requested file, with average file size being 5MB. We observe that the greedy system policy is able to support an intensity



(a) Spatial Load Profile 1

(b) Spatial Load Profile 2

(c) Spatial Load Profile 3

(d) Spatial Load Profile 4

(e) Spatial Load Profile 5

(f) Three Base Stations: Homogeneous Load

Fig. 9: ISAP: Delay performance

of arrivals up to 10% more than the greedy user policy while the proposed policy ISAP as well as the projection based optimized static policy support a further increase of 50% over even the greedy system policy. The results indicate that ISAP performs well even in a multiple base station network.

VIII. Performance Sensitivity

Channel Model: We use parameters that model cellular base stations in an urban environment. We simulate a system consisting of two base stations 2800 meters apart, and compare the performance of the schemes presented earlier using a path loss exponent 3.5, and a cell-edge signal to noise ratio of 10 dB. The data rate at which users are served is calculated using Shannon's capacity formula, after a 6dB backoff is applied to the perceived SINR. Fig. 10 shows the simulated delay performance when load split between the base stations is varied from 0.5 (even division of load) to 0.1, similar to Fig. 2. The results again show that the optimal load division depends on the intensity of the offered load, and is not balanced but significantly asymmetric. Fig. 11 compares the delay performance of the optimized static policy and ISAP to the dynamic schemes described earlier. The proposed schemes significantly outperform the dynamic schemes. The mean delay under the greedy system scheme, for example, is over 50 times the mean delay under the optimized static scheme at high loads. These results demonstrate that the performance trends observed earlier do not depend on the particular parameters chosen for the propagation model, but hold for more realistic ones as well, and are a consequence of the dynamics introduced by inter-cell interference.

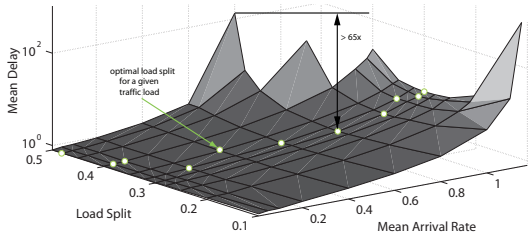


Fig. 10: Optimal load split thresholds

Long tailed file size distributions: In the process of determining the optimized static threshold, we still assume that file sizes are exponentially distributed. We assume that the

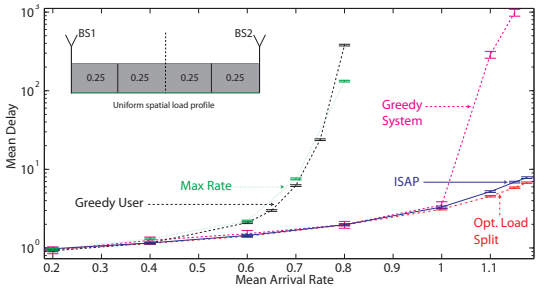


Fig. 11: Sensitivity to channel model

users' file size requirements are log normally distributed with mean 5 MB, and variance 12.276×10^6. The performance of the various schemes under a spatially homogeneous user distribution is shown in Fig. 12. The relative performance of the different schemes is very similar to the case of exponential file sizes. While the insensitivity of mean queue length to file size distribution under processor sharing is a well known result in the case of a single queue, the results indicate that mean delay might not be significantly impacted by file size distribution even in a system with interacting servers. The optimized static policy and the policy developed above, ISAP, result in the best performance. The optimization procedure and the proposed adaptive policy appear to be robust to variations in the distribution of users' file size requirements.

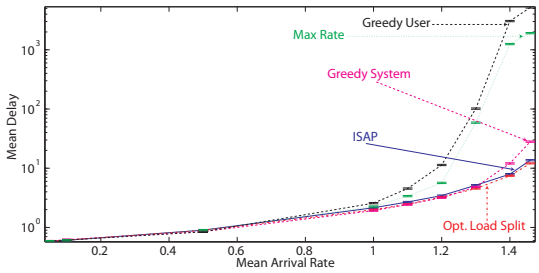


Fig. 12: Sensitivity to file size distribution

IX. A Braess's Paradox Like Phenomenon

Generally, we expect that the addition of resources in the form of extra base stations would only result in improved user performance. Even in an interference-dominated scenario, an optimal user association policy could, in the worst case, ensure that no users are assigned to the additional base station resulting in no change in performance. Here, we present a scenario where the addition of a fourth base station results in worse user performance on average under some user association disciplines. This can be viewed as a type of Braess's paradox [27] for an infrastructure based wireless network.

The scenario under consideration: We consider a service area that is a circle of radius 100 m that is served by three base stations that are distributed evenly around the circumference (the solid base stations in the inset in Fig. 13). We investigate the effect on user-perceived delay performance of adding a fourth base station (the dashed BS in Fig. 13) at the center of the circle. Note that the base station at the center is located at a position that is proximate to the users that are farthest away from the three original base stations. We simulate a non-homogeneous spatial user distribution. 50% of the user arrivals are distributed homogeneously within the service area while the rest are restricted to arrive uniformly within a ring of width 4m located halfway between the center of the service area and the edge. While the scenario considered here might not be representative of real-world conditions, the results serve to demonstrate the counter-intuitive nature of the user association problem in the highly coupled, non-linear, interference-dominated setting and emphasize the importance of carefully designing a policy that accounts for these factors.

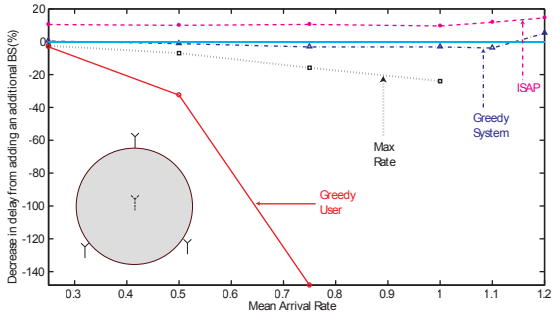


Fig. 13: Evaluating changes in performance from adding a fourth base station

Fig. 13 exhibits the percentage reduction in mean user-perceived file transfer delay when a base station is added to the center of the service area under each of the user association policies discussed in this paper. The max rate and greedy user policies demonstrate a significant increase in the mean delays when the extra base station is added. In the case of the greedy user policy mean delays are more than doubled, while under the max rate policy, mean delays increase by more than 20% at high user arrival rates. These policies do not explicitly account for the effects of interference, and can make decisions that result in an overall increase in the level of inter-cell interference resulting in poor performance. Under the greedy system policy user performance slightly worsens at low user arrival rates, but shows marginal improvement at the highest system loads. Under ISAP, user performance does improve. Mean delays are observed to be lower by about 10% at low user arrival rates, with further improvement as system load increases. This demonstrates that the addition of an extra base station can indeed improve performance even in this interference-dominated setting if appropriate user association policies are used. Note that once again ISAP outperforms the other user association polices when only three base stations are used and the gap between policies increases as a result of this phenomenon.

X. Conclusion

We considered a user-base station association problem in wireless networks serving dynamic loads and thus coupled through interference and proposed a methodology to bound and optimize performance of such systems. For the one and two dimensional models considered, the performance gain

from optimized static policies is substantial, even outperforming natural greedy user and system dynamic policies. The load-balancing static policy was shown to be very poor, showing that the critical aspect is inducing asymmetry in the load, even when the network and loads are symmetric. We presented a novel interference-sensitive, adaptive user association policy for multi-base station networks. Our simulation results demonstrated that our proposed policies perform consistently well under all spatial loads and are robust to variations in file size distributions and large-scale propagation parameters. The performance of the conventional dynamic policies was found to vary dramatically with the load distribution, and no one policy performed consistently well. This work suggests the possibility that substantial gains might be achieved if network functions (see e.g., Sec. I) coupled through interference (or otherwise) are optimized for dynamic loads.

APPENDIX

The following definitions provide a characterization of the stochastic ordering relationship between two process, and will be used in the proof of Lemma 3.1.

Definition 10.1 ([28]): Let $\mathbf{l}$, $\mathbf{m} \in \mathbb{R}^n$, and let $l_{[1]} \geq \cdots \geq l_{[n]}$ denote the components of $\mathbf{l}$ arranged in descending order.

$$\mathbf{l} \prec_w \mathbf{m} \text{ if } \sum_{i=1}^{k} l_{[i]} \leq \sum_{i=1}^{k} m_{[i]},\ k = 1, \ldots, n$$

The vector $\mathbf{l}$ is then said to be *weakly majorized* by $\mathbf{m}$.

Definition 10.2 ([29]): Let $\mathbf{L}$, $\mathbf{M}$ be random vectors taking values in $\mathbb{R}^n$. $\mathbf{L}$ is *stochastically weak-majorized* by $\mathbf{M}$, written $\mathbf{L} \prec_w^{\text{st}} \mathbf{M}$, if there exist random vectors $\tilde{\mathbf{L}}$ and $\tilde{\mathbf{M}}$ taking values in $\mathbb{R}^n$ with the same probability laws as $\mathbf{L}$ and $\mathbf{M}$ respectively, with $\tilde{\mathbf{L}} \prec_w \tilde{\mathbf{M}}$ a.s.

***Proof of Lemma 3.1*:** We will demonstrate that the policy π_b, which is obtained from π_a by exchanging service regions $\mathcal{R}_1$ and $\mathcal{R}_2$ between the base stations, obtains a lower (or equal) mean delay, see Section III. This is shown by constructing a pair of coupled processes $\tilde{\mathbf{U}}^{\pi_a}(t)$ and $\tilde{\mathbf{U}}^{\pi_b}(t)$, such that

$$\tilde{\mathcal{U}}_1^{\pi_b}(t) \subseteq \tilde{\mathcal{U}}_1^{\pi_a}(t) \text{ and } \tilde{\mathcal{U}}_2^{\pi_b}(t) \subseteq \tilde{\mathcal{U}}_2^{\pi_a}(t), \tag{1}$$

and such that $\tilde{\mathbf{U}}^{\pi_a}(t) \sim \mathbf{U}^{\pi_a}(t)$ and $\tilde{\mathbf{U}}^{\pi_b}(t) \sim \mathbf{U}^{\pi_b}(t)$. It follows that associated queue length processes $\tilde{\mathbf{Q}}^{\pi_\mathbf{a}}(\mathbf{t})$ and $\tilde{\mathbf{Q}}^{\pi_\mathbf{b}}(\mathbf{t})$ satisfy similar properties with containment replaced with an inequality. By standard arguments, see [29], this construction suffices to show that $\mathbf{Q}^{\pi_\mathbf{b}}(t)$ is *stochastically weak-majorized* by $\mathbf{Q}^{\pi_\mathbf{a}}(t)$. As $t \to \infty$ this implies π_b achieves a lower (or equal) mean queue length, and thus, by Little's Law, a lower (or equal) mean delay.

Note that the arrival rates associated with the exchanged service regions are equal so the arrival rate to each base station under the two policies are the same, i.e., $\lambda_1 = \lambda(\mathcal{X}_1^{\pi_a}) = \lambda(\mathcal{X}_1^{\pi_b})$ and $\lambda_2 = \lambda(\mathcal{X}_2^{\pi_a}) = \lambda(\mathcal{X}_2^{\pi_b})$. We couple arrivals of the two processes $\tilde{\mathbf{U}}^{\pi_a}(t)$ and $\tilde{\mathbf{U}}^{\pi_b}(t)$, as generated by a common Poisson process with intensity $\lambda_1 + \lambda_2$. For convenience, we index user requests based on arrival times (including those in the system at $t = 0$), i.e., $1, 2, \ldots$ While arrival times for users to the two systems are identical, their locations may not be, whence we let $x_i^{\pi_a}$ and $x_i^{\pi_b}$ denote the locations of the i^{th} request under policy π_a and π_b respectively.

Suppose $x \in \tilde{\mathcal{U}}_1^{\pi_a}(t)$ then let $c_x^{\pi_a}(t)$ be the capacity to the user under policy π_a at time t taking into account the state of the neighboring base station. Since users share capacity via processor sharing, effective service rate to users at locations x and y under the two policies is given by $\mu^{\pi_a}(t,x) = \frac{c_x^{\pi_a}(t)}{Q^{\pi_a}_{\beta^{\pi_a}(x)}(t)}$. and $\mu^{\pi_b}(t,y) = \frac{c_y^{\pi_b}(t)}{Q^{\pi_b}_{\beta^{\pi_b}(y)}(t)}$. So the departure rate of users from BS1 under policy π_a is given by

$$\mu_1^{\pi_a}(t) = \sum_{x \in \tilde{\mathcal{U}}_1^{\pi_a}(t)} \mu^{\pi_a}(t,x).$$

We define the overall departure rates $\mu_2^{\pi_a}(t)$, $\mu_1^{\pi_b}(t)$, and $\mu_2^{\pi_b}(t)$ analogously.

Let $\tilde{\mathbf{U}}^{\pi_a}(0) = \tilde{\mathbf{U}}^{\pi_b}(0)$ so (1) holds at time $t = 0$. Our construction will be such that if (1) holds at some time t then it is satisfied after the next arrival/departure, while maintaining marginal dynamics that are consistent with systems associated with policies π_a and π_b. Although the two systems see the same overall arrival rates they may see different overall departure rates. In our construction we let

$$\nu(t) = \lambda_1 + \lambda_2 + \max(\mu_1^{\pi_a}(t),\, \mu_1^{\pi_b}(t)) + \max(\mu_2^{\pi_a}(t),\, \mu_2^{\pi_b}(t))$$

denote the current rate of events for the *coupled processes* and allow fictitious events to ensure the marginal system processes have the correct dynamics. Let the time at which the next event occurs be t' and z be a realization of a random variable Z, which is uniformly distributed on $[0, \nu(t)]$. The coupled process events are constructed as follows:

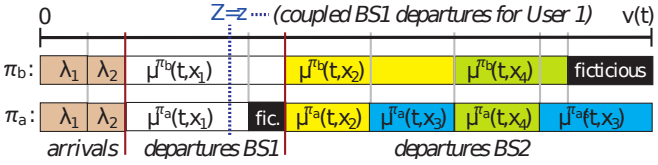


Fig. 14: Example coupling construction for arrivals/departures based on realization of Z.

Arrivals: If $0 \leq z \leq \lambda_1$, the next event is an arrival, say of user n , to BS1 under both policies. We let random variables $X_n^{\pi_a}$ and $X_n^{\pi_b}$ denote the position of this user under policies π_a and π_b respectively. The distribution $X_n^{\pi_a}$ is given by $\mathbf{P}(X_n^{\pi_a} \in A) = \frac{\lambda(A)}{\lambda_1}$, for a measurable set $A \subseteq \mathcal{X}^{\pi_a}$. The position of the user under policy π_b is identical, except if $X_n^{\pi_a} \in \mathcal{R}_1$. In this case, the user's location falls within $\mathcal{R}_2$ with a distribution $\mathbf{P}(X_n^{\pi_b} \in B | X_n^{\pi_a} \in \mathcal{R}_1) = \frac{\lambda(B)}{\lambda(\mathcal{R}_2)}$, where $B \subseteq \mathcal{R}_2$. The states of the processes are updated accordingly. If $\lambda_1 \leq z \leq \lambda_1 + \lambda_2$, the next event is an arrival to BS2 under both policies, with the user's location generated analogously to the above. In either case, arrivals to BS1 or BS2 occurs simultaneously for both policies, so (1) holds at time t'. Also under the above construction the spatial distribution of Poisson arrivals is maintained.

Departures: If $\lambda_1 + \lambda_2 \leq z \leq \lambda_1 + \lambda_2 + \max(\mu_1^{\pi_a}(t),\, \mu_1^{\pi_b}(t))$, the event is a potential departure from BS1. Consider any user k such that $x_k^{\pi_b} \in \tilde{\mathcal{U}}_1^{\pi_b}(t)$. Since (1) holds, user k is also in the

system under policy π_a, i.e., $x_k^{\pi_a} \in \tilde{\mathcal{U}}_1^{\pi_a}(t)$. Since (1) holds there are only three cases to consider:

1. $\tilde{\mathcal{U}}_2^{\pi_b}(t) = \tilde{\mathcal{U}}_2^{\pi_a}(t) = \emptyset$: BS2 is idle under both policies. If $x_k^{\pi_a} = x_k^{\pi_b}$, $c^{\pi_b}_{x_k^{\pi_b}}(t) = c^{\pi_a}_{x_k^{\pi_a}}(t)$. Otherwise, $x_k^{\pi_a} \in R_1$ and $x_k^{\pi_b} \in R_2$, so Fact 3.1 implies $c^{\pi_b}_{x_k^{\pi_b}}(t) \geq c^{\pi_a}_{x_k^{\pi_a}}(t)$.

2. $\tilde{\mathcal{U}}_2^{\pi_b}(t) \neq \emptyset, \tilde{\mathcal{U}}_2^{\pi_a}(t) \neq \emptyset$: BS2 is transmitting under both policies, and, as in the previous case, we can argue that $c^{\pi_b}_{x_k^{\pi_b}}(t) \geq c^{\pi_a}_{x_k^{\pi_a}}(t)$.

3. $\tilde{\mathcal{U}}_2^{\pi_b}(t) = \emptyset, \tilde{\mathcal{U}}_2^{\pi_a}(t) \neq \emptyset$: In this case, users in BS1 see no interference under policy π_b while they see interference from BS2 under policy π_a. Combining our conclusion in case 1 with the fact that the data rate at which users can be served is an increasing function of the received signal to interference plus noise ratio, we see that $c^{\pi_b}_{x_k^{\pi_b}}(t) \geq c^{\pi_a}_{x_k^{\pi_a}}(t)$.

Also, by assumption $\tilde{Q}_1^{\pi_b}(t) \leq \tilde{Q}_1^{\pi_a}(t)$, thus $\mu^{\pi_b}(t, x_k^{\pi_b}) \geq \mu^{\pi_a}(t, x_k^{\pi_a})$. This permits us to couple User k's departure such that if it leaves under policy π_a, it also leaves under policy π_b. To see this, consider Fig. 14 where $[0, \nu(t)]$ has been subdivided based on the arrival rates and service rates of the users in the system under the two policies. If a user is present in both systems then a set of length $\mu^{\pi_a}(t, x_k^{\pi_a})$ for policy π_a is contained within one of length $\mu^{\pi_b}(t, x_k^{\pi_b})$ for policy π_b. If the user has already left the system under policy π_b, the corresponding set for policy π_a can be arranged arbitrarily (need not be contiguous) within $[0, \nu(t)]$. Unused intervals correspond to dummy events. Which departures (if any) occur for the two systems depend on which sets contain z. However, clearly a departure of User k from BS1 under policy π_a results in the same under policy π_b unless it has already left the system, and (1) still hold at time t'. If $(\lambda_1 + \lambda_2 + \max(\mu_1^{\pi_a}(t), \mu_1^{\pi_b}(t))) \leq z$, the event is a potential departure from BS2, and is treated analogously to departures from BS1.

Since relationship (1) holds after any future event, by induction the relationship holds for all times in the future. It immediately follows that

$$\begin{aligned} \tilde{Q}_1^{\pi_a}(t) &\geq \tilde{Q}_1^{\pi_b}(t), \\ \tilde{Q}_2^{\pi_a}(t) &\geq \tilde{Q}_2^{\pi_b}(t). \end{aligned}$$

Thus, we have shown that $\mathbf{Q}^{\pi_b}(t)$ is *stochastically weak-majorized* by $\mathbf{Q}^{\pi_a}(t)$. As $t \to \infty$ this implies π_b achieves a lower (or equal) mean queue length, and thus, by Little's Law, a lower (or equal) mean delay.

REFERENCES

[1] S. Borst, N. Hegde, and A. Proutiere, "Capacity of wireless data networks with intra- and inter-cell mobility," in *INFOCOM*, 2006.

[2] S. Borst, "User-level performance of channel-aware scheduling in wireless data networks," in *INFOCOM*, 2003.

[3] T. Bonald, S. Borst, and A. Proutiere, "Inter-cell coordination in wireless data networks," *European Trans. on Telecommunications*, vol. 17, pp. 303–312, 2006.

[4] S. K. Das, S. K. Sen, and R. Jayaram, "A dynamic load balancing strategy for channel assignment using selective borrowing in cellular mobile environment," *Wireless Networks*, vol. 3, no. 5, pp. 333–47, 1997.

[5] G. Bianchi and I. Tinnirello, "Improving load balancing mechanisms in wireless packet networks," in *IEEE ICC*, vol. 2, 2002.

[6] E. Yanmaz, O. K. Tonguz, and R. Rajkumar, "Is there an optimum dynamic load balancing scheme?" in *IEEE Globecom*, vol. 1, 2005.

[7] K. Navaie and H. Yanikomeroglu, "Downlink joint base-station assignment and packet scheduling algorithm for cellular CDMA/TDMA networks," in *IEEE ICC*, vol. 9, 2006, pp. 4339–44.

[8] S. Das, H. Viswanathan, and G. Rittenhouse, "Dynamic load balancing through coordinated scheduling in packet data systems," in *INFOCOM*, 2003.

[9] A. Sang, X. Wang, M. Madihian, and R. D. Gitlin, "Coordinated load balancing, handoff/cell-site selection, and scheduling in multi-cell packet data systems," in *ACM MobiCom*, 2004, pp. 302–14.

[10] E. Altman, A. Kumar, C. K. Singh, and R. Sundaresan, "Spatial SINR games combining base station placement and mobile association," in *IEEE INFOCOM*, 2009, pp. 19–25.

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

[12] S. Borst, I. Saniee, and A. Whiting, "Distributed dynamic load balancing in wireless networks," in *ITC*, 2007, pp. 1024–37.

[13] A. Zemlianov and G. de Veciana, "Load balancing of best effort traffic in wirel. sys. supporting end nodes with dual mode capabilities," in *CISS*, 2005.

[14] T. Bonald, S. Borst, N. Hegde, and A. Proutiere, "Wireless data performance in multicell scenarios," in *SIGMETRICS*, 2004.

[15] S. Borst, N. Hegde, and A. Proutiere, "Interacting queues with server selection and coordinated scheduling - application to cellular data networks," *Annals of Operations Research*, vol. 170, no. 1, pp. 59–78, September 2009.

[16] M. Jonckheere, "Stability of two interfering processors with load balancing," in *Third International Conference on Performance Evaluation Methodologies and Tools*, 2008.

[17] S. C. Borst, "Optimal probabilistic allocation of customer types to servers," in *ACM SIGMETRICS*, 1995, pp. 116–25.

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

[19] F. Guillemin and D. Pinchon, "Analysis of generalized processor sharing systems with two classes of customers and exponential services," *Journal of Applied Probability*, vol. 41, no. 3, pp. 832–858, 2004.

[20] S. Borst., O. Boxma, and P. Jelenkovic, "Coupled processors with regularly varying service times," in *IEEE INFOCOM*, vol. 1, 2000, pp. 157–64.

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

[23] T. S. Rappaport, *Wireless Communications: Principles and Practice*. Prentice Hall, 2002.

[24] B. Rengarajan and G. de Veciana, "Architecture and abstractions for environment and traffic aware system-level coordination of wireless networks: The downlink case," in *INFOCOM*, 2008, pp. 502–10.

[25] ——, "User association to optimize flow level performance in wireless systems with dynamic interference," in *NETCOOP*, November 2009.

[26] B. Rengarajan, C. Caramanis, and G. de Veciana, "Analyzing queueing systems with coupled processors through semidefinite programming," http://users.ece.utexas.edu/~gustavo/papers/SdpCoupledQs.pdf, 2008.

[27] D. Braess, "Uber ein paradoxon aus der verkehrsplannung," *Unternehmensforschung*, vol. 12, pp. 258–268, 1968.

[28] A. Marshall and I. Olkin, *Inequalities: Theory of Majorization and its Applications*. New York: Academic Press, 1979.

[29] D. Stoyan, *Comparison Methods for Queues and Other Stochastic Models*. New York: John Wiley, 1983.

Balaji Rengarajan received his Ph.D. and M.S. in electrical engineering from the University of Texas at Austin in 2009 and 2004 respectively, and his B.E. in Electronics and Communication from the University of Madras in 2002. He was the recipient of a 2003 Texas Telecommunications Engineering Consortium (TxTEC) graduate fellowship. He is currently a researcher at IMDEA networks, Madrid.

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.