User Association to Optimize Flow Level Performance in Wireless Systems with Dynamic Interference ⋆

Balaji Rengarajan and Gustavo de Veciana

[1] IMDEA Networks {balaji.rengarajan@gmail.com}
[2] University of Texas at Austin {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). Systematic simulations establish that our optimized static policy substantially outperforms various dynamic policies and that these results are robust to changes in file size distributions, channel parameters, and spatial load distributions.
Key words: User association, flow-level performance, coupled queues, semidefinite programming

1 Introduction

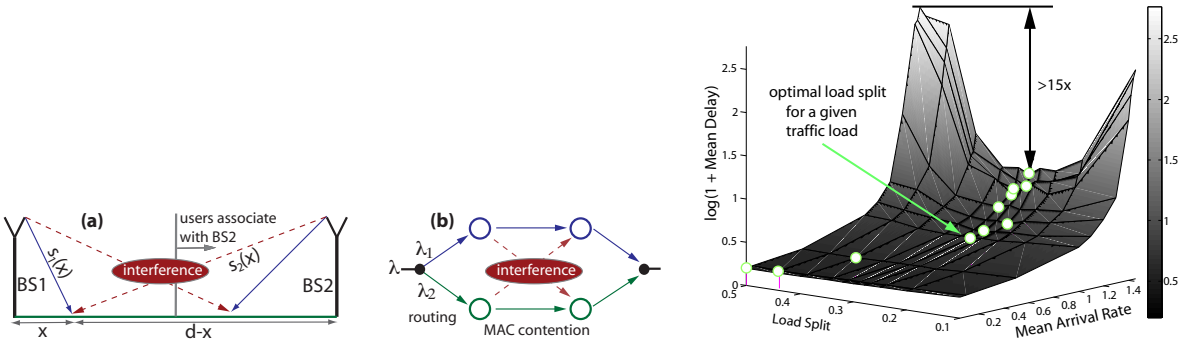
The high demand for wireless capacity and the increasing volume of traffic mandates the efficient use of available radio resources. Wireless capacity can be substantially enhanced by reusing the entire frequency spectrum at every transmitter instead of sacrificing individual peak and overall system capacity 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.

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.

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?

⋆ This research was supported by NSF Award CNS-0917067, and Intel Research Council Grant, and performed in part under the auspices of IMDEA Redes in Madrid.



(a) User Association (b) MAC/routing (c) User Association: Impact of load split threshold on delay performance.

Fig. 1. Examples of network functions impacted by dynamic coupling

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. Fig. 1c 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. This result is surprising, and reveals 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.

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.

The case of dynamic traffic, with the associated bursty interference, has not been extensively studied. In [10], 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 [11], but under the assumption that transmissions are orthogonal. The impact of dynamic interference was also demonstrated in [12], wherein the problem of load balancing in a hybrid wireless local area/wide area network was studied using approximations proposed in [13].

The stability region of a dynamic system with interacting servers under load balancing strategies was examined in [14]. 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 [15]. 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 flow level performance, i.e., the actual file transfer delays experienced by users. We obtain bounds on the mean delay experienced by users in coupled systems and use these bounds to design user association policies that optimize flow level performance.

Our contributions: We study a static load allocation scheme that takes into account the long term spatial load being served by the base stations and attempts to minimize the average file transfer delay perceived by users. In addition to short-term, unpredictable variations in the load caused by individual user arrivals and departures, there are predictable long-term variations in the aggregate traffic load depending on the day-of-week, hour-of-day, etc. [16, 11]. The proposed scheme is adapted to these long-term traffic patterns, and does not depend on the instantaneous loads or short-term variations. Such a scheme is potentially simpler than dynamic schemes which require knowledge of the instantaneous loads being served, but might be sensitive to errors in the long term traffic estimates. 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 or good approximations were available for general systems with 3 or more coupled queues.

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 demonstrate through extensive simulations that the proposed policy 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.

Organization of paper: The system model is described in detail in Sec. 2. The optimal static association policy is characterized in Sec. 3, while Sec. 4 explores the impact of asymmetric static association policies. The methodology used to pick the optimal static policy is presented in Sec. 5. Simulation results comparing the performance of the static policy to various dynamic strategies is presented in Sec. 6, and the effect of non-homogeneous spatial loads is explored in Sec. 6.4. The sensitivity of delay performance to file size distributions and system and channel parameters is considered in Sec. 6.5, and Sec. 7 concludes the paper.

2 System Model

In Secs. 3-4, 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 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 and only asymptotic results and closed form expressions in the case of some simple work-conserving scenarios with two coupled queues are known. Even the problem of characterizing the stability of coupled queues which was addressed in [22] is difficult, and one has to employ numerical methods.

2.1 Simulation Model

In the bulk of the simulation results, we consider two base stations located 500m apart with users arriving according to a Poisson process.The three base station network studied in Sec. 6.3 consists of three facing sectors in a hexagonal layout of base stations with cell radius 250m, with users again arriving according to a Poisson process.In Secs. 3-6, where we develop and study the semidefinite programming based methodology, we assume that the user distribution 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. 6.4. 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. Shadowing, and fading are not considered in these preliminary results. 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. 6.5 where a system with a higher path loss exponent, and cell-edge SNR of 10 dB is simulated.

3 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 1. *Consider the two base station model defined in Sec. 2. 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. 2. 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.

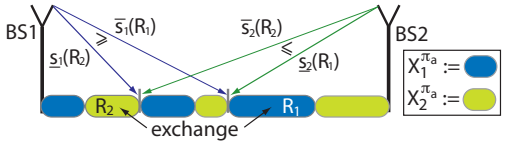


Fig. 2. A sub-optimal load allocation policy.

Fact 1 *Under the assumptions on $\mathcal{R}_1$ and $\mathcal{R}_2$ in Lemma 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 1. *For the two base station model defined in Sec. 2, 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 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 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.

4 Optimal Threshold Trends

As a consequence of Theorem 1, we need only consider threshold-based static allocation policies. Fig. 1c 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. As noted earlier, 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 were 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.

5 Optimizing the Threshold

In this section, we propose an approximation methodology for optimizing static load allocation policies for the wireless network model in Sec. 2, 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 $\lambda_n = \lambda(\mathcal{X}_n^\pi)$. Several technical challenges will be addressed. 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. Second, we propose an approach to upper/lower bound the performance for the 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 the accuracy of the proposed methodology is excellent. We also note that the approach is applicable to a broader set of problems with coupled queues or dynamic interference.

5.1 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. 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 $\vec{\delta}^i$.

The incremental time 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. 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 chain with transition rate bounded by $\eta = \sum_{n=1}^N \lambda_n + N\mu^*$, so it can be uniformized. This will be of use in the sequel. 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

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

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. When the system is stable [25, 26], there is a stationary distribution for $(\vec{Q}, \vec{X})$ such that

$$\vec{Q} \stackrel{d}{=} g(\vec{Q}, \vec{X}) := \vec{Q} + \vec{X} \tag{1}$$

where $\stackrel{d}{=}$ denotes equality in distribution.

5.2 Performance Bounds

Below, we describe our approach to bounding the system's mean sum-queue length. The approach extends the work of [27] to coupled queuing systems and can also be used to bound other performance metrics, see [28] for more details. Bounds on the mean queue lengths in turn translate to bounds on the mean delay via Little's Law.

Let ψ denote a joint distribution for $(\vec{Q}, \vec{X})$ on $\mathcal{S} = \mathcal{S}_{\vec{Q}} \times \mathcal{S}_{\vec{X}} \subseteq \mathbb{Z}_+^N \times \mathbb{Z}_+^N$ satisfying (1) and with marginals $\psi_{\vec{Q}}$ and $\psi_{\vec{X}}$. Eq. (1) can in this case be rewritten as

$$\psi_{\vec{Q}} = \psi g^{-1}. \tag{2}$$

We partition $\mathcal{S}_{\vec{Q}}$ into 2^N regions where the *same* set of queues are non-zero, i.e., $\mathcal{S}_{\vec{\delta}^i} := \{\vec{q} : \delta_n^i = \mathbf{1}(q_n > 0), n = 1, \ldots, N\}$ for $i = 1, \ldots, 2^N$. Let $\psi^{\vec{\delta}^i}$ and $\psi_{\vec{Q}}^{\vec{\delta}^i}$, $\psi_{\vec{X}}^{\vec{\delta}^i}$ be the conditional distributions for $(\vec{Q}, \vec{X})$ and its marginals given $\vec{Q} \in \mathcal{S}_{\vec{\delta}^i}$. Note that for all states in $\mathcal{S}_{\vec{\delta}^i}$ the queues share the same service rates, so follows $\vec{Q}$ is conditionally independent of $\vec{X}$ given $\vec{Q} \in \mathcal{S}_{\vec{\delta}^i}$, i.e.,

$$\psi^{\vec{\delta}^i} = \psi_{\vec{Q}}^{\vec{\delta}^i} \psi_{\vec{X}}^{\vec{\delta}^i}, \; i = 1, \ldots, 2^N. \tag{3}$$

We shall use multi-index notation in formulating our bounds. For $\vec{\alpha} \in \mathbb{Z}_+^N$ and $\vec{Y} \in \mathbb{R}^N$ we let $\vec{Y}^{\vec{\alpha}}$ denote the term $Y_1^{\alpha_1} \ldots Y_N^{\alpha_N}$, and let $|\vec{\alpha}| = \sum_{n=1}^N \alpha_n$. For $r \in \mathbb{N}$ we define

$$m_{\vec{\delta}^i}^{\vec{\beta}} = \mathbf{E}_{\psi_{\vec{X}}^{\vec{\delta}^i}} \left[\vec{X}^{\vec{\beta}} \right], \; |\vec{\beta}| \leq 2r \text{ and } i = 1, \ldots, 2^N. \tag{4}$$

Given the transition probabilities on each region $\mathcal{S}_{\vec{\delta}^i}$ these can be easily computed. Bounds on mean sum-queue length can be obtained by optimizing distributions ψ satisfying the following constraints:

Problem 1. Given $\mathcal{S}, \mathcal{S}_{\vec{Q}}$ and $\mathcal{S}_{\vec{X}}$ solve:

$$\begin{aligned} \sup / \inf_{\psi} \;\; & \mathbf{E}_{\psi_{\vec{Q}}} \left[\sum_{n=1}^N Q_n \right] \\ \text{s.t.} \;\; & (2), (3), (4), \\ & \mathbf{E}_{\psi}[1] = \mathbf{E}_{\psi_Q}[1] = \mathbf{E}_{\psi_X}[1] = 1, \qquad (5) \\ & \psi \in \mathbb{M}(\mathcal{S}), \psi_Q \in \mathbb{M}(\mathcal{S}_Q), \psi_X \in \mathbb{M}(\mathcal{S}_X). \qquad (6) \end{aligned}$$

Here, $\mathbb{M}(\mathcal{S})$, $\mathbb{M}(\mathcal{S}_{\vec{Q}})$, and $\mathbb{M}(\mathcal{S}_{\vec{X}})$ are sets of positive Borel measures supported on $\mathcal{S}$, $\mathcal{S}_Q$, and $\mathcal{S}_X$ respectively, and (5) ensures they are probability measures. The parameter r controls the degree of accuracy of such bounds [29]. As $r \to \infty$ the distribution of $\vec{X}$ is specified exactly, in turn uniquely determining the distributions of $\vec{Q}$ and and $(\vec{Q},\vec{X})$.

To allow numerical computation, we further relax Problem 1 based on joint moments of degree no higher than $2r$. For all $\vec{\alpha},\vec{\beta}$ such that $|\vec{\alpha}|+|\vec{\beta}| \leq 2r$ and $k = 1,\ldots,2^N$ we define decision variables:

$$x_k^{\vec{\alpha}\vec{\beta}} := \mathbf{E}[\vec{Q}^{\vec{\alpha}}\vec{X}^{\vec{\beta}} | \vec{Q} \in \mathcal{S}_{\vec{\delta}^k}]\ \mathbf{P}(\vec{Q} \in \mathcal{S}_{\vec{\delta}^k}).$$

Note that the mean sum queue length can now be expressed as $\sum_{n=1}^{N}\sum_{k=1}^{2^N} x_k^{\vec{e}_n,0}$, where $\vec{e}_n$ is the unit vector with a 1 for queue n.

Eq. (2), or equivalently (1), implies equality for all moments on the left and right hand sides. Note that

$$g(\vec{Q},\vec{X})^{\vec{\alpha}} = (\vec{Q}+\vec{X})^{\vec{\alpha}} = \sum_{|\vec{\gamma}_1|+|\vec{\gamma}_2|\leq\alpha} g_{\vec{\alpha}}^{(\vec{\gamma}_1,\vec{\gamma}_2)} \vec{Q}^{\vec{\gamma}_1}\vec{X}^{\vec{\gamma}_2},$$

with Binomial coefficients $\{g_{\vec{\alpha}}^{(\vec{\gamma}_1,\vec{\gamma}_2)}\}$. So, we can relax the distributional constraint (2) to a constraint on moments of order no higher than $2r$ giving

$$\sum_{k=1}^{2^N} x_k^{\vec{\alpha},0} = \sum_{k=1}^{2^N} \sum_{|\vec{\gamma}_1|+|\vec{\gamma}_2|\leq\alpha} g_{\vec{\alpha}}^{(\vec{\gamma}_1,\vec{\gamma}_2)} x_k^{\vec{\gamma}_1\vec{\gamma}_1},\ \forall |\vec{\alpha}| \leq 2r. \tag{7}$$

We also relax constraint (3) by equating the moments of the product distribution to the products of the moments

$$x_k^{\vec{\alpha}\vec{\beta}} = m_{\delta_k}^{\vec{\beta}} x_k^{\vec{\alpha},0},\ \forall |\vec{\alpha}|+|\vec{\beta}| \leq 2r,\ k = 1,\ldots,2^N, \tag{8}$$

which also subsumes (4). Constraint (5) can be directly expressed as:

$$\sum_{k=1}^{2^N} x_k^{0,0} = 1. \tag{9}$$

Finally we need to ensure constraints (6) are satisfied by the moments/decision variables. Let $x_k = \{x_k^{\vec{\alpha}\vec{\beta}},\ |\vec{\alpha}|+|\vec{\beta}| \leq 2r\}$, a moment sequence associated with the set $\mathcal{S}_{\vec{\delta}^k}$ for $k = 1,\ldots,2^N$. We denote the cone of moments supported on $\mathcal{S}_k = \mathcal{S}_{\vec{\delta}^k} \times \mathcal{S}_X$ by $\mathcal{M}_{2r}(\mathcal{S}_k)$, and it's closure by $\overline{\mathcal{M}_{2r}(\mathcal{S}_k)}$. So (6) translates to a moment constraint $x_k \in \overline{\mathcal{M}_{2r}(\mathcal{S}_k)},\ \forall k = 1,\ldots,2^N$. Such constraints can in turn be expressed as certain matrices being positive semidefinite. In general, a necessary condition for the sequence $y = \{y^{\vec{\alpha}\vec{\beta}},\ |\vec{\alpha}|+|\vec{\beta}| \leq 2r\}$ to be a valid truncated sequence of moments is that the associated moment matrix, denoted $\mathbf{M}_r(y)$, be positive semidefinite, i.e., $\mathbf{M}_r(y) \succeq 0$ – see [27, 29] for details. Such constraints must be satisfied by the truncated moment sequences, i.e.,

$$\mathbf{M}_r(x_k) \succeq 0,\ \ k = 1,\ldots,2^N. \tag{10}$$

Additionally, note that $\mathcal{S}_{\vec{\delta}^k}$ can be specified by an intersection of linear inequalities and so can $\mathcal{S}_k$, i.e.,

$$\mathcal{S}_k = \cap_{h\in H_k}\{h(\vec{q},\vec{x}) \geq 0\},$$

where H_k denotes a set linear functions defining $\mathcal{S}_k$. Clearly $\{x_k\}$ should also be a valid truncated moment sequence when restricted to each half plane $\{h(\vec{q},\vec{x}) \geq 0\}$. Again it can be shown that a necessary condition for this to be the case is that an associated (localizing) moment matrix, denoted $\mathbf{M}_{r-1}(h,x_k)$, depending on the coefficients of the hyperplane h and x_k be positive semidefinite, see [27, 29] for details. The corresponding set constraints is given by

$$\mathbf{M}_{r-1}(h,x_k) \succeq 0,\ \forall h \in H_k, k = 1,\ldots,2^N. \tag{11}$$

Substituting our relaxed moment and semidefinite constraints into Problem 1 we obtain :

Problem 2. Given $m_{\vec{\delta}^i}^{\vec{\beta}},\ |\vec{\beta}| \leq 2r$ and $i = 1,\ldots,2^N$, the moments of X, solve:

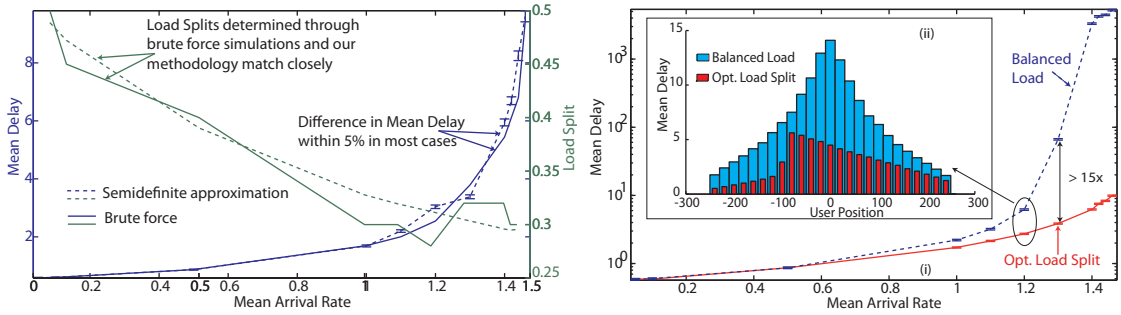
$$\sup / \inf_{\{x_k|k=1,\ldots,2^N\}} \sum_{k=1}^{2^N}\sum_{n=1}^{N} x_k^{\vec{e}_n,0}$$
$$\text{s.t. } (7),(8),(9),(10),(11).$$

Solving this semidefinite optimization problem also yields tighter bounds as r is increased, i.e., more information about X is used, yet at an increased complexity. We shall see that small r suffice in the sequel.

5.3 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 1 shows the optimal static load allocation policy is a simple threshold. So for any threshold, Problem 2 can be solved 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. 6.3.

Fig. 3a exhibits the computed approximate optimal thresholds versus those obtained via brute force simulation for our two base station model. Semidefinite optimization problems associated with relaxations of order 2 were solved using [30] and [31] to determine the necessary bounds. 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 however provides the flexibility to address complex traffic loads as well as systems with a larger number of base stations.



(a) Goodness of optimized thresholds. (b) Optimized policy vs. load balancing.

Fig. 3. Performance of the optimized static user association policy.

6 Performance Comparison

6.1 Comparing Static Policies

Fig. 3b-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. Fig. 3b-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.

6.2 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 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.

Fig. 4a 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

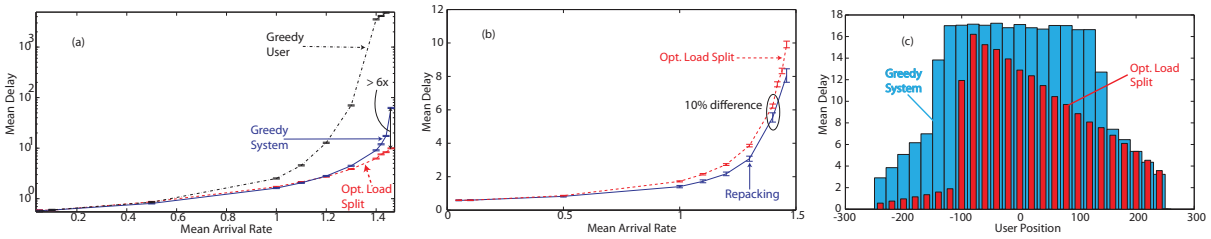


Fig. 4. Mean delay under optimized policy vs. (a) greedy schemes (log scale); (b) repacking scheme (linear scale); (c) greedy system in terms of spatial distribution.

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. 4b (linear scale) is the best, but indeed very close to the optimal static policy.

Fig. 4c 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.

6.3 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. 2.1, 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.

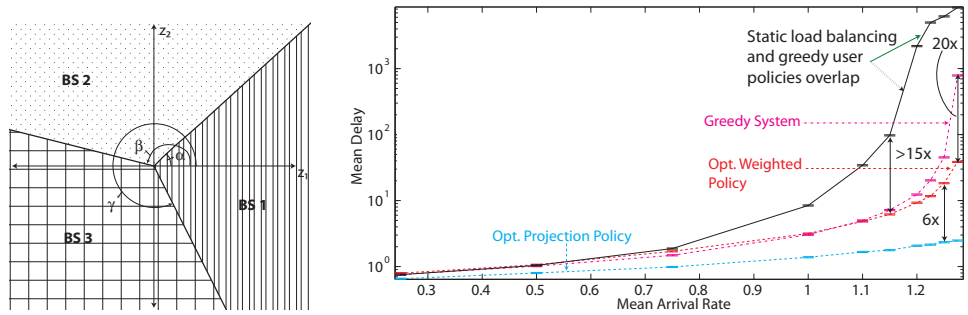
Weighted signal strengths The first family of policies we consider is parametrized by base station weights. Each base station is assigned a weight and a user is associated with the base station that offers the maximum weighted received signal strength. The weight associated with one of the base stations is set to 1, and a simple gradient descent is used to determine weights for the remaining base stations. The bounding methodology described in Sec. 5.2 is used to approximate the mean delay at each step of the gradient descent algorithm.

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. 6.3 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.

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. 5a. 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 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



(a) Projecting to the two dimensional hyperplane

(b) Mean delay performance

Fig. 5. Three base station network

chosen. Thus, each iteration lowers the overall mean delay experienced by users in the system, ensuring that the optimization procedure converges.

Fig. 5b 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.

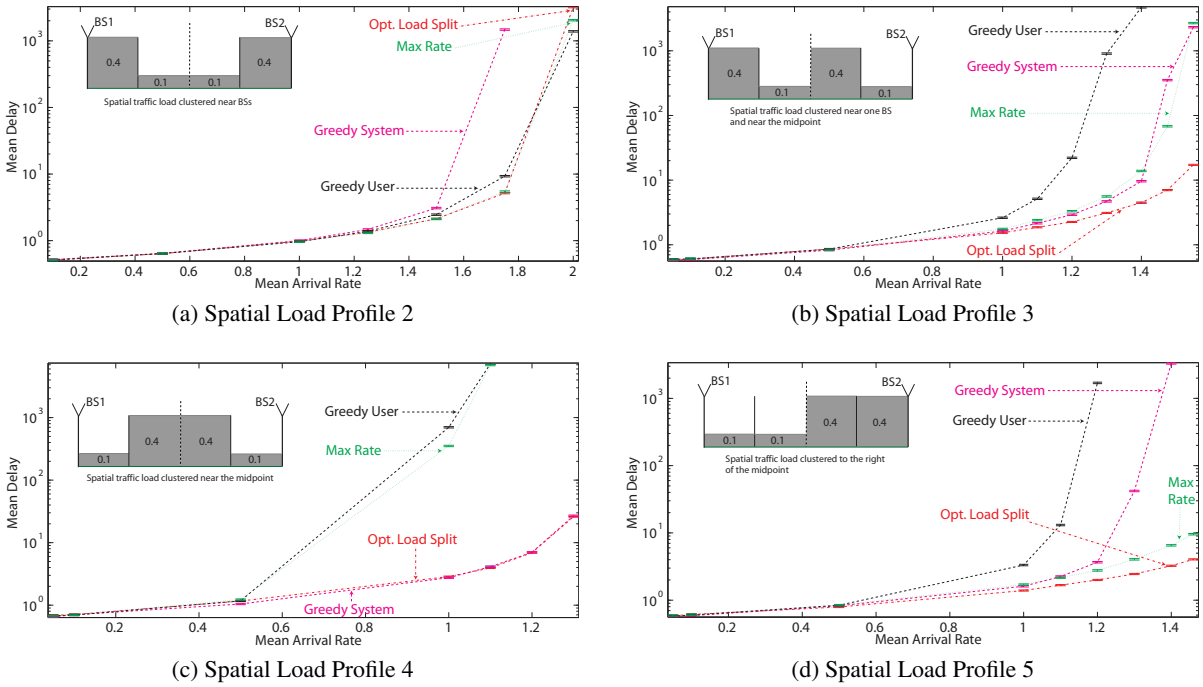
6.4 Heterogeneous Spatial Traffic Loads

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 for the two base station scenario, as shown in Fig. 6a-6d. 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 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 performs consistently well under all spatial load profiles, and performs as well as or outperforms all the dynamic policies. This demonstrates its robustness to spatially heterogeneous traffic loads. Under load profile 3, for example, the optimized static policy outperforms all the other schemes by a wide margin. None of the other schemes perform well under all profiles, while our proposed scheme is able to infer the nature of the spatial load and adapt to it. 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.

6.5 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. 7a demonstrates that the optimized static policy significantly outperforms 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 channel model used and are a consequence of the dynamics introduced by inter-cell interference.



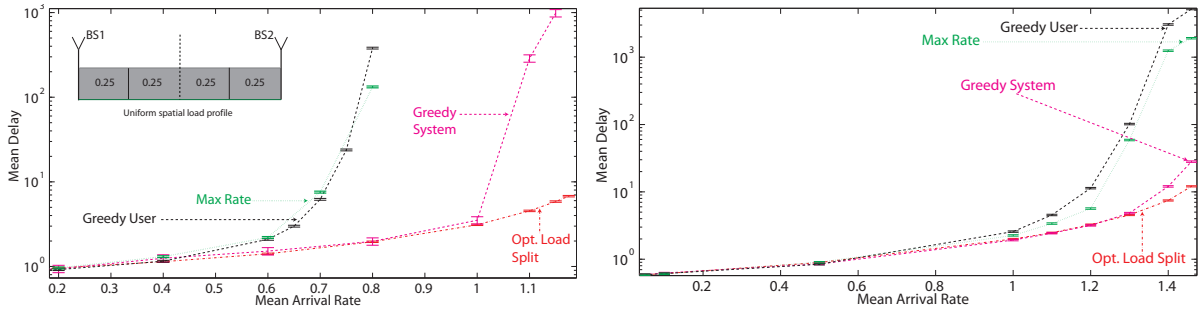
(a) Spatial Load Profile 2

(b) Spatial Load Profile 3

(c) Spatial Load Profile 4

(d) Spatial Load Profile 5

Fig. 6. Delay performance under spatially heterogeneous traffic



(a) Channel model

(b) File size distribution

Fig. 7. Performance Sensitivity

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 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. 7b. The relative performance of the different schemes is very similar to the case of exponential file sizes. The optimized static policy again results in the best performance, and appears to be robust to variations in the distribution of users' file size requirements.

7 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. Our simulation results demonstrated that our proposed policy performs consistently well under all spatial loads and is robust to variations in file size distributions and channel 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. 1) coupled through interference (or otherwise) are optimized for dynamic loads.

References

1. Borst, S., Hegde, N., Proutiere, A.: Capacity of wireless data networks with intra- and inter-cell mobility. In: INFOCOM. (2006)
2. Borst, S.: User-level performance of channel-aware scheduling in wireless data networks. In: INFOCOM. (2003)
3. Bonald, T., Borst, S., Proutiere, A.: Inter-cell coordination in wireless data networks. European Transactions on Telecommunications **17** (2006) 303–312
4. Das, S.K., et al.: A dynamic load balancing strategy for channel assignment using selective borrowing in cellular mobile environment. Wirel. Netw. **3**(5) (1997) 333–47
5. Bianchi, G., Tinnirello, I.: Improving load balancing mechanisms in wireless packet networks. In: IEEE ICC. Volume 2. (2002)
6. Yanmaz, E., et al.: Is there an optimum dynamic load balancing scheme? In: IEEE GTC. Volume 1. (2005)
7. Navaie, K., Yanikomeroglu, H.: Downlink joint base-station assignment and packet scheduling algorithm for cellular CDMA/TDMA networks. In: IEEE ICC. Volume 9. (2006) 4339–44
8. Das, S., et al.: Dynamic load balancing through coordinated scheduling in packet data systems. In: INFOCOM. (2003)
9. Sang, A., et al.: Coordinated load balancing, handoff/cell-site selection, and scheduling in multi-cell packet data systems. In: ACM MobiCom. (2004) 302–14
10. Bonald, T., et al.: Inter-cell scheduling in wireless data networks. In: European Wireless Conference. (2005)
11. Borst, S., Saniee, I., Whiting, A.: Distributed dynamic load balancing in wireless networks. In: ITC. (2007) 1024–37
12. Zemlianov, A., et al.: Load balancing of best effort traffic in wirel. sys. supporting end nodes with dual mode capabilities. In: CISS. (2005)
13. Bonald, T., et al.: Wireless data performance in multicell scenarios. In: SIGMETRICS. (2004)
14. Borst, S., Hegde, N., Proutiere, A.: Interacting queues with server selection and coordinated scheduling - application to cellular data networks. Annals of Operations Research **170**(1) (September 2009) 59–78
15. Jonckheere, M.: Stability of two interfering processors with load balancing. In: Third IInternational Conference on Performance Evaluation Methodologies and Tools. (2008)
16. Borst, S., et al.: Dynamic optimization in future cellular networks. Bell Labs Technical Journal **10**(2) (2005) 99–119
17. Borst, S.C.: Optimal probabilistic allocation of customer types to servers. In: ACM SIGMETRICS. (1995) 116–25
18. Fayolle, G., Lasnogorodski, R.: Two coupled processors: The reduction to a Riemann–Hilbert problem. Wahrscheinlichkeitstheorie (3) (Jan. 1979) 1–27
19. Guillemin, F., Pinchon, D.: Analysis of generalized processor sharing systems with two classes of cus and exponential services. J. Appl. Prob. **41**(3) (2004) 832–858
20. Borst., S., et al.: Coupled processors with regularly varying service times. In: IEEE INFOCOM. Volume 1. (2000) 157–64
21. Borst, S., et al.: The asymptotic workload behavior of two coupled queues. Queueing Systems **43**(1-2) (January 2003) 81–102
22. Borst, S., Jonckheere, M., Leskelä, L.: Stability of parallel queueing systems with coupled service rates. Discrete Event Dynamic Systems **18**(4) (2008) 447–472
23. Rappaport, T.S.: Wireless Communications: Principles and Practice. Prentice Hall (2002)
24. Rengarajan, B., de Veciana, G.: Architecture and abstractions for environment and traffic aware system-level coordination of wireless networks: The downlink case. In: INFOCOM. (2008) 502–10
25. Borovkov, A., Foss, S.: Stochastically recursive sequences and their generalizations. Siberian Adv. in Math. **2**(1) (1992) 16–81
26. Loynes, R.: The stability of a queue with non-independent inter-arrival and service times. Proc. Cambr. Phil. Soc. **58** (1962) 497–520
27. Bertsimas, D., Natarajan, K.: A semidefinite optimization approach to the steady-state analysis of queueing systems. Queueing Syst. Theory Appl. **56**(1) (2007) 27–39
28. Rengarajan, B., Caramanis, C., de Veciana, G.: Analyzing queueing systems with coupled processors through semidefinite programming. `http://users.ece.utexas.edu/~gustavo/papers/SdpCoupledQs.pdf` (2008)
29. Lasserre, J.: Bounds on measures satisfying moment conditions. Annals of Applied Probability **12** (2002) 1114–1137
30. Henrion, D., Lasserre, J.: Gloptipoly: global optimization over polynomials with matlab and sedumi. ACM Transactions on Mathematical Software **29**(2) (2003) 165–194
31. Sturm, J.F., the Advanced Optimization Laboratory at McMaster University: Sedumi version 1.1r3 (2006) See sedumi.mcmaster.ca.
32. Marshall, A., Olkin, I.: Inequalities: Theory of Majorization and its Applications. New York: Academic Press (1979)
33. Stoyan, D.: Comparison Methods for Queues and Other Stochastic Models. New York: John Wiley (1983)

Appendix

The following definitions provide a characterization of the stochastic ordering relationship between two process, and will be used in the proof of Lemma 1.

Definition 1 ([32]). *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 2 ([33]). *Let* $\mathbf{L}, \mathbf{M}$ *be random vectors taking values in* $\mathbb{R}^n$. $\mathbf{L}$ *is* stochastically weak-majorized *by* $\mathbf{M}$*, written* $\mathbf{L} \prec_w^{st} \mathbf{M}$*, 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 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 3. 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

$$\tilde{\mathcal{U}}_1^{\pi_b}(t) \subseteq \tilde{\mathcal{U}}_1^{\pi_a}(t) \text{ and } \tilde{\mathcal{U}}_2^{\pi_b}(t) \subseteq \tilde{\mathcal{U}}_2^{\pi_a}(t), \tag{12}$$

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 [33], 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 (12) holds at time $t = 0$. Our construction will be such that if (12) 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:

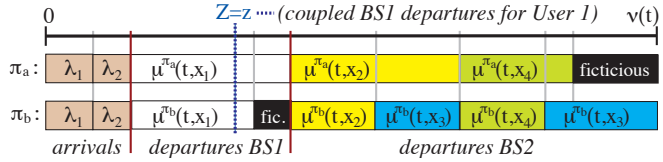


Fig. 8. Example coupling construction for arrivals/deparartures based on realization of Z.

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 (12) 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 (12) 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 (12) 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 $x_k^{\pi_a} = x_k^{\pi_b}$, $c_{x_k^{\pi_b}}^{\pi_b}(t) = c_{x_k^{\pi_a}}^{\pi_a}(t)$. Otherwise, $x_k^{\pi_a} \in R_1$ and $x_k^{\pi_b} \in R_2$, so Fact 1 implies $c_{x_k^{\pi_b}}^{\pi_b}(t) \geq c_{x_k^{\pi_a}}^{\pi_a}(t)$.

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 $c_{x_k^{\pi_b}}^{\pi_b}(t) \geq c_{x_k^{\pi_a}}^{\pi_a}(t)$.

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 $c_{x_k^{\pi_b}}^{\pi_b}(t) \geq c_{x_k^{\pi_a}}^{\pi_a}(t)$.

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. 8 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_a, the corresponding set for policy π_b 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 (12) 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 (12) holds after any future event, by induction the relationship holds for all times in the future. We show that the following relationship hold at any given time

1. $\tilde{Q}_1^P(t) \geq \tilde{Q}_1^{P_E}(t)$ and $\tilde{Q}_2^P(t) \geq \tilde{Q}_2^{P_E}(t)$
2. Corresponding to every user attached to BS1 under policy P_E, there exists a user attached to BS1 under policy P, that is served at lower rates both when BS2 is idle and active.
3. Corresponding to every user attached to BS2 under policy P_E, there exists a user attached to BS2 under policy P, that is served at lower rates both when BS1 is idle and active.

Now, consider a sequence of user arrivals and departures resulting from a static load allocation policy that associates users in region r_1 with base station 1, and users in region r_2 with base station 2. We construct an alternate sample path based on this sequence of user arrivals. User arrivals in region r_1 are moved to instead arrive in region r_2 while still being served by base station 1, and user arrivals in region r_2 are moved to r_1 and served by base station 2. All other user arrivals are unchanged. Since the probability of user arrivals in region r_1 is equal to the probability of user arrivals in r_2, and there are no correlations between user arrivals, the constructed sequence of user arrivals is also representative of the arrival process.

The user queues associated with the two base stations evolve identically until a user arrives in either region r_1 or r_2. As proved previously, this user is served at a higher rate in the alternate sample path. All other users currently in the queues are served at exactly the same rates as in the original sample path as the queue lengths of the two queues are identical. As a result the user that arrived to one of the regions r_1 or r_2 is served and leaves the system earlier in the alternate sample path. This results in all the other remaining users in the system being served at the same or greater rate. Thus, all users in the alternate sample path perceive delays that are less than or equal to those in the original sample path.

Since both systems are ergodic, the sample mean of the user delays in both sample paths converge eventually to the expected values, and the expected user delay in the alternate sample path has to be lower or equal to that in the original. Note that this alternate sample path corresponds to a policy which associates users in region r_2 with base station 1, and users in r_1 to base station 2.