Traffic-Driven Spectrum Allocation in Heterogeneous Networks

Binnan Zhuang, Dongning Guo, and Michael L. Honig
EECS Dept, Northwestern University, Evanston, IL 60208, USA

July 14, 2014

Abstract

This paper studies the optimization of resource allocation in wireless heterogeneous networks (HetNets) on a relatively slow timescale. Unlike most previous work on physical layer resource allocation, this work considers traffic dynamics and uses the average packet sojourn time rather than the sum rate and the outage probability as the performance metric. The approach addresses user QoS in HetNets with many small cells, where large traffic variations in overlapping cells cause complicated interference conditions. To be specific, in a HetNet cluster with \( K \) base terminal stations, each with a queue for dynamic traffic arrivals, we determine the optimal partition of the spectrum into \( 2^K - 1 \) segments, corresponding to all possible spectrum sharing combinations. The \( K \) queues are coupled since their instantaneous service rates depend on the interference from other cells. Two spectrum allocation schemes are obtained taking an upper bound and a refined approximation of the average packet sojourn time, respectively. The optimization problem based on the upper bound is convex, and it is proved that the optimal allocation divides the spectrum into at most \( K \) segments. The refined approximation lies between upper and lower bounds, and the corresponding optimization problem admits an efficient suboptimal solution. Simulation results show that the refined allocation outperforms both the orthogonal and full-spectrum-reuse allocations under all traffic conditions. The improvement is significant in the heavy traffic regime, in which case the two proposed allocation schemes perform similarly.
I. INTRODUCTION

Dense deployment of small cells is a major means to address the scarcity of spectrum resources for future cellular networks [2]. By reducing the coverage of each micro/pico base terminal stations (BTS) and frequency reuse, the capacity of a network can be significantly increased to accommodate increasing traffic demands. With overlapping cells of all sizes, a heterogenous network (HetNet) often operates in the interference limited regime. Small cells may also lead to more pronounced traffic variations. Hence traditional frequency reuse and cell planning can be ineffective. Resource allocation according to dynamic traffic then becomes highly desirable.

One conventional spectrum allocation scheme is fractional frequency reuse (FFR). Dynamic FFR can improve the sum-rate and total throughput of a network through effectively mitigating inter-cell interference [3]–[5]. However these results are based on the assumption that traffic is backlogged at all BTS’s. In a dense HetNet with dynamic traffic, the sum-rate is not a good indicator of user QoS. In this work, we focus on the average packet delay under offered traffic as the QoS metric, which better characterizes the user experience in a network. With dynamic traffic, each BTS alternates between service and vacation periods, which introduces time-varying inter-cell interference. A recent study in [6] points out that the backlogged traffic assumption, i.e., ‘always transmitting/interfering,’ exaggerates the contribution of each BTS to inter-cell interference in the small-cell scenario. Here we propose a traffic-dependent interference model to address this issue. The optimization problems in [3]–[5] are considered on fast timescales with instantaneous information exchange. Since traffic variations happen on a slower timescale in practice (compared to the timescale of scheduling) and there are limitations on the rate at which a central controller can acquire traffic and service information from BTS’s, the spectrum allocation problem is solved by a central controller on a slow timescale, as in [7], [8], using aggregate traffic and service information.

The objective of the optimization is to improve user QoS through physical layer resource adaptation, as the spatial load in the HetNet changes. The key is to connect the spectrum and power resources in the physical layer to the QoS in the network layer. This is accomplished using the service rates of the queues at all BTS’s as the link. To be specific, in a cluster of \( K \) cells, an allocation divides the available spectrum into \( 2^K - 1 \) segments, corresponding to all possible sharing combinations. A given spectrum allocation across cells induces an interference
pattern and corresponding spectral efficiency within each segment of the spectrum according to Shannon’s capacity formula. The spectral efficiencies along with the bandwidths of the segments allocated to a BTS determine the service rate of the BTS. The average packet delay at a given BTS is in turn determined by the service rates and the packet arrival rates. Thus an optimization problem is formulated with the average packet sojourn time as the objective, the bandwidths of the segments as desired variables, and the service rates as intermediate variables. The resource allocation problem is a joint physical layer and network layer optimization problem integrating results from both information theory and queueing theory.

The queueing dynamics in the proposed system can be modeled by interactive queues. Since it is difficult to obtain a closed-form steady-state distribution of interactive queues, we resort to approximation techniques. In Section IV an upper bound on the average packet sojourn time is obtained by serving each queue independently at a conservative rate, which can be sustained regardless of the state of the other queues. The problem of optimizing the upper bound turns out to be convex. It is proved that the optimal allocation divides the spectrum into at most $K$ segments (in lieu of $2^K - 1$ possible segments). Numerical results show that the conservative allocation greatly reduces the average packet delay in the heavy traffic regime compared to both the orthogonal and the full-spectrum-reuse allocations. However, the conservative allocation is suboptimal in light and moderate traffic regimes due to the pessimistic assumption on the service rates.

The refined approximation is developed by studying the $K$ dimensional continuous-time Markov chain (CTMC) of the interactive queueing system. A product form distribution is used to approximate the true steady-state distribution. According to the numerical results, the refined approximation is close to the actual delay in all traffic regimes. In simulations the optimal refined allocation always achieves the minimum delay compared to the orthogonal, full-spectrum-reuse and conservative allocations. The performance gain is observed mainly because each BTS is dynamically driven by the spatial loads to allocate a sufficient amount of spectrum to serve its own traffic demands, while leaving enough spectrum for adjacent BTS’s.

The remainder of this paper is organized as follows. Some related work is reviewed in Section II. The system model is introduced in Section III and the spectrum allocation problems based on conservative and refined approximations are presented in Section IV and V respectively. Numerical results are given in Section VI and conclusions are given in Section VII.
II. RELATED WORK

The flow-level performance of a similar multi-cell system with inter-cell interference was studied in [9] using the bounds developed in [10] for queues with time-varying capacity. The model in [9] corresponds to full-spectrum-reuse, hence spectrum allocation is not considered. Moreover, the focus of [9] is to evaluate the delay for fixed traffic arrival rates and service rates, while the goal of this paper is to improve system-wide delay through resource allocation.

The network throughput of a multi-cell system was studied in [11] by investigating the stability region of the corresponding interactive queueing system. It is shown that interference avoidance through inter-cell scheduling provides large throughput gains. In contrast, the resource allocation schemes in this paper not only achieve the throughput region of the network for all possible spectrum allocations (as shown in Section V-D), but also reduce the average delay for any given traffic load inside the throughput region.

Traffic- and topology-aware resource allocation was also considered in [12]. Two major differences between [12] and this work are: Resource allocation in [12] is performed in the time domain using iteratively update scheduling policy and BTS utilizations; and the queuing policy of [12] is processor sharing in contrast to the ‘first in first out’ (FIFO) policy assumed in this paper. Both [12] and this work indicate the importance of adapting to the traffic and topology.

III. SYSTEM MODEL

We consider downlink data transmission in a HetNet with \( K \) arbitrarily deployed BTS’s. Denote the set of all BTS’s as \( S = \{1, \ldots, K\} \). The \( K \) BTS’s collectively share one unit of bandwidth to serve their respective user equipments (UEs). A central controller determines which part of the spectrum is allocated to each BTS. Assuming the frequency resources are interchangeable, the problem is equivalent to deciding the bandwidth shared by each subset of BTS’s, denoted by a \( 2^K \)-tuple: \( x = (x(B))_{B \in 2^S} \), where \( 2^S = \{B \mid B \subseteq S\} \) is the power set of \( S \) containing all combinations of BTS’s (including the empty one), and \( x(B) \in [0, 1] \) is the fraction of spectrum shared by BTS’s in set \( B \). Clearly, \( \sum_{B \in 2^S} x(B) = 1 \), and any efficient allocation would set \( x(\emptyset) = 0 \). For example, if \( K = 2 \), the three variables \( x(\{1\}) \), \( x(\{2\}) \) and \( x(\{1, 2\}) \) denote the amount of spectrum allocated to BTS 1 and BTS 2, exclusively, and that shared by both BTS’s.
We next specify the relation between spectrum allocation and the service rate at each BTS. The actual service rate in each cell depends on its own spectrum usage as well as the interference from other cells. To characterize the interaction among multiple BTS’s, we define $A$ as the set of BTS’s that are transmitting data to their UEs at a given time. It is assumed that BTS $i$ applies flat power spectral density (PSD) $p_i$ over the allocated spectrum whenever it actively transmits. The spectral efficiency of BTS $i$ on a segment of spectrum denoted by $s_i(C)$, is a function of the set $C$, which consists of all BTS’s actively sharing the segment. The service rate in cell $i$ when the active set is $A$ is thus given by:

$$r_i(A) = \sum_{B \in 2^S} s_i(B \cap A)x(B).$$  \hspace{1cm} (1)

The intersection of sets in (1) is because among all BTS’s in $B$, only those in $B \cap A$ are transmitting.

The spectral efficiency function should satisfy $s_i(A) = 0$ if $i \not\in A$, i.e., if BTS $i$ is not active, then its spectral efficiency is zero. Moreover, if $A \subseteq B$, then $s_i(A) \geq s_i(B)$, i.e., the efficiency decreases with more active BTS’s. For concreteness in obtaining numerical results, we will assume the spectral efficiency of BTS $i$ with active sharing set $A$ is given by:

$$s_i(A) = 1(i \in A) \times \log \left( 1 + \frac{p_i}{I_i(A) + n_i} \right) \text{ (bits/second/Hz)},$$ \hspace{1cm} (2)

where $1(i \in A) = 1$ if $i \in A$ and $1(i \in A) = 0$ otherwise, $I_i(A)$ is the constant interference PSD from BTS’s in $A$, and $n_i$ is the noise PSD. The actual values of $s_i(A)$’s depend on the transmit PSD at each BTS, the path-loss model and network topology. The UEs of a cell are assumed to be at the same point to simplify the interference model. The model can be refined to account for different locations of the UEs in a cell by allowing multiple types of users with different service rates, which would require more involved queueing analysis and is left for future work. Since the optimization will be performed on a slow timescale, the spectral efficiencies are assumed known by the central controller a priori.

User service requests are modeled as packet arrivals at each BTS following a Poisson process with rate $\lambda_i$ at BTS $i$. All packet lengths are independent identically distributed according to the exponential distribution with unit mean. Given $\lambda_i$ and $s_i(A)$ for all $i \in S$ and $A \in 2^S$, the objective is to minimize the average mean packet sojourn time (aka the delay) by optimizing $x$. To evaluate the flow-level performance, we assume user requests within a cell are processed
according to the FIFO criterion. However the results in this paper apply to all queueing disciplines that are work-conserving, non-anticipating and non-preemptive as defined in [13].

The $K$ BTS’s form a network of $K$ interactive queues, where the instantaneous service rate of each queue depends on the status of the other queues at the same time. Such an interactive queueing system is also referred to as a *coupled-processors* model. In the special case of two coupled queues, finding the joint steady-state distribution can be formulated as a Riemann-Hilbert problem [14]. The same two-dimensional Markov process has been studied in [15], where the steady-state distribution can be represented as an infinite series of product of powers. Two coupled processors with generally distributed service times have been studied in [16], which shows the joint workload distribution can be determined by solving a boundary value problem. These results are difficult to use for numerical computation. Also, few results exist for more than two coupled queues. Here we provide two approximations of the mean packet sojourn time as functions of the spectrum allocation. The first is an upper bound on the sojourn time, where the queues become decoupled M/M/1 queues by assuming a conservative service rate. Although the approximation is pessimistic, the effect on packet delay due to load variation is preserved. The second approximation is more refined by including inter-cell interactions at the cost of increased computational complexity.

**IV. Delay Upper Bound Conservative Spectrum Allocation**

In this section, we introduce a first approximation of the mean packet sojourn time. The key is to assume each BTS always transmits at the worst-case rate, which is achievable regardless of other BTSs’ states. This assumption is equivalent to assuming other BTS’s are always backlogged and interfering. Hence each BTS $i$ serves its UEs with constant rate $r_i(S)$, which is given by [1] with $A = S$. The $K$ queues become independent M/M/1 queues. Throughout this section, we use $r_i$ to denote $r_i(S)$ for simplicity. The average packet sojourn time at BTS $i$ takes a simple form [17]:

$$ t_i = \frac{1}{r_i - \lambda_i}. $$

(3)
A. The Optimization Problem

The spectrum allocation problem based on the conservative approximation (3) is formulated as:

\[
\text{minimize } \{x(B), B \subseteq S\} \sum_{j=1}^{K} \frac{\lambda_j}{\sum_{j=1}^{K} \lambda_j} \sum_{i=1}^{\lambda_i} \lambda_i - \lambda_i \]  
\text{(P1a)}
\]

subject to

\[r_i = \sum_{B \subseteq S} s_i(B)x(B), \forall i \in S \]  
\text{(P1b)}
\[
r_i > \lambda_i, \forall i \in S \]  
\text{(P1c)}
\[
x(B) \geq 0, \forall B \in 2^S \]  
\text{(P1d)}
\[
\sum_{B \subseteq S} x(B) = 1. \]  
\text{(P1e)}

The objective (P1a) is the weighted average delay of the entire network, where \(\lambda_i / \sum_{j=1}^{K} \lambda_j\) is the fraction of total traffic served by BTS \(i\). The constraints (P1c) guarantee the stability of all queues. Problem (P1) is a convex optimization problem, because by treating \(x(B)\)'s and \(r_i\)'s as the optimization variables, all the constraints are linear, and the objective is a linear combination of convex functions as \(t(r) = 1 / (r - \lambda)\) is convex in \(r\) on \((\lambda, \infty)\).

Since the objective function (P1a) is strictly convex and bounded from below, according to [18] the optimization problem (P1) has a unique global minimum when feasible. Due to the special structure of the conservative spectrum allocation problem, the optimal solution has the following property.

\textbf{Theorem 1:} In the optimal solution of the \(K\)-BTS conservative spectrum allocation problem, the spectrum is divided into at most \(K\) segments:

\[
\left\{ B \mid x(B) > 0, B \in 2^S \right\} \leq K. \]  
\text{(4)}

\textbf{Proof:} Denote the rate vector \(r\) and spectral efficiency vector \(s(B)\) as \(r = [r_1, \ldots, r_K]\) and \(s(B) = [s_1(B), \ldots, s_K(B)]\). According to (P1b) to (P1c), \(r \in \mathbb{R}_+^K\) is a convex combination of the \(2^K - 1\) points \(\{s(B), B \in 2^S\}\), with coefficients \(\{x(B), B \in 2^S\}\), i.e., \(r = \sum_{B \in 2^S} s(B)x(B)\). In other words, any \(r\) given by (P1b) is in the convex hull of \(\{s(B), B \in 2^S\}\). By Carathéodory’s Theorem [19], \(r\) lies in a \(d\)-simplex with vertices in \(\{s(B), B \in 2^S\}\) and \(d \leq K\), i.e., \(r\) can be written as a convex combination of \(x(B), B \in 2^S\) with at most \(K + 1\) nonzero \(x(B)\)'s. This holds for any \(r\) satisfying (P1b) to (P1e). Furthermore, the \(r^*\) corresponding to the optimal solution
to (P1) must be Pareto optimal in terms of the rate allocation, i.e., there is no spectrum allocation \(\bm{x}\), such that \(r^*_i \leq \sum_{B \subseteq S} s(B)x(B), \ \forall i \in S\) and at least one of the inequality is strict. This is because any spectrum allocation that could increase the rate at any BTS without decreasing the rates at other BTS’s would also decrease the objective (P1a). Hence \(r^*\) cannot be an interior point of the \(d\)-simplex, and must lie on some \(m\)-face of the \(d\)-simplex with \(m < d \leq K\). Therefore \(r^*\) can be written as a convex combination of \(m + 1 \leq K\) nonzero \(x(B)\)’s.

**Corollary 1:** In the optimal solution of the \(K\)-BTS conservative spectrum allocation problem, for any subset \(M \subseteq S\) of the BTS’s with \(m\) BTS’s, the spectrum exclusively used by those BTS’s in \(M\) is divided into at most \(m\) segments:

\[
\left| \{B \mid x(B) > 0, B \in 2^M\} \right| \leq m.
\] (5)

**Proof:** If the bandwidths of the spectrum segments not in \(2^M\) are fixed at their optimal values, then (P1) becomes an optimization problem of variables \(x(B), B \in 2^M\), the service rates at BTS’s not in \(M\) are fixed and the service rates for the \(m\) BTS’s in \(M\) are convex combinations of \(x(B), B \in 2^M\) plus a constant vector in \(R^m\). Therefore Corollary 1 can be proved using a similar argument as in the proof of Theorem 1.

\[\square\]

**B. An Efficient Algorithm**

The structure of the optimal solution given by Theorem 1 suggests the possibility of solving (P1) more efficiently. Originally we need to determine the sizes of \(2^K - 1\) spectrum segments. The computational complexity is polynomial in \(2^K\) using a standard convex optimization algorithm. By Theorem 1 we only need to decide the sizes of the \(K\) nonzero segments. The difficulty is to decide which set of \(K\) segments out of the \(2^K - 1\) possibilities. Algorithm 1 is an iterative algorithm for solving the \(K\)-BTS spectrum allocation problem.

1) **Initialization:** The algorithm needs to start at a feasible point. A standard method is to solve a modified version of optimization problem (P1) by replacing the objective function (P1a) with a constant. By introducing slack variables, the resulting optimization problem can be transformed to a linear program in standard form, which can be solved using the simplex method [20]. Although the worst-case complexity of this method is \(O(2^K)\), for most practical problems the complexity is usually polynomial in \(K\). According to the properties of basic feasible solution to a linear program [20], the solution (initial point for (P1)) will have at most \(K + 1\) nonzero \(x(B)\)’s.
Algorithm 1 Iterative algorithm for solving the conservative spectrum allocation problem

**INPUT:** $\lambda_i$ and $s_i(B)$ for all $i \in S$ and $B \in 2^S$.

**OUTPUT:** $x(B)$ for all $B \in 2^S$.

**Initialization:** Find a feasible solution $x_0(B)$ by solving (P1) with constant objective. $N = \{B \mid x_0(B) > 0\}$, $N' = \emptyset$, $N^+ = 2^S$.

**while** $N^+ \not\subset N'$ **do**

1. $N' = N$;

2. Find $x(B)$ by solving (P1) starting from $x_0(B)$ with the additional constraints, $x(B) = 0$, $\forall B / \in N$;

3. Compute the partial derivatives of the objective function (P1a) with respect to all $x(B)$’s, $\Delta x(B) = -\sum_{i \in B} \frac{\lambda_i s_i(B)}{(r_i - \lambda_i)^2}$;

4. $N^+ = \{B$ for $K$ smallest $\Delta x(B)\}$, $N = N \cup N^+$, $x_0(B) = x(B)$.

**end while**

2) **Algorithm Description:** Starting from a feasible point $(x_0(B))_{B \in 2^S}$, let $N$ be the candidate set, which initially includes the indices of those nonzero spectrum segments of the initial point. In each iteration, the algorithm finds the optimal solution within the candidate set $N$. After each iteration, the partial derivatives with respect to all $x(B)$’s are calculated (including those not in $N$). The $K$ segments with the $K$ smallest derivatives are added to the candidate set. (The number of variables added to the candidate set may be less than $K$, since some of the $B$’s of the $K$ smallest derivatives may already be in $N$.) The algorithm iterates with the solution found in the last iteration as the new initial point and the expanded candidate set. The algorithm stops when the candidate set stops growing. At the end of each iteration, if the solution is not optimal, there must be some $B$’s outside the candidate set that have smaller partial derivatives. Since we only add more $B$’s to the candidate set with each iteration, in the worst-case, the candidate set will eventually include all $2^K - 1$ variables. Hence the proposed algorithm is guaranteed to converge to the global optimum.

3) **Performance:** The algorithm is more efficient when starting at an initial point with fewer nonzero spectrum segments. Therefore we can always use the full-spectrum-reuse allocation, $x(S) = 1$, as an initial point if it is feasible. Even if it is not, the solution obtained by the
initialization method has no more than $K + 1$ nonzero spectrum segments. Examples of delay versus number of iterations are shown in Fig. 1 with $K = 7$ and different traffic loads. In the simulation, Algorithm 1 starts with the full-spectrum-reuse allocation. The figure shows that the algorithm converges within a few iterations.

The allocation scheme developed in this section is suboptimal because it assumes conservative transmission rates based on the worst-case interference level. As we shall see, the gap between this conservative solution and the refined solution introduced in Section V becomes significant in the case of light traffic.

V. REFINED SPECTRUM ALLOCATION

In this section, we derive a refined approximation for the delay taking into account the interactions due to dynamic interference. The proposed approximation takes into consideration these complicated interactions, but make simplifications to yield analytical results. The best known upper and lower bounds on the average delay are achieved by considering these interactions in the limiting regimes to be described in Section V-C. It is shown in Section V-D that the refined approximation is between existing upper and lower bounds.

A. A Refined Approximation of the Average Delay

The original $K$-dimensional CTMC can be described as follows. Let $l = [l_1, \ldots, l_K]$ be the state of the CTMC, where $l_i \geq 0$ is the number of packets in queue $i$ (including the packet
Fig. 2. The continuous time Markov Chain for the 2-BTS interactive queueing model.

being served, if any). Let \( A_l = \{ i \mid l_i > 0 \} \) denote the corresponding set of active BTS’s. The transition rate from state \( l \) to state \( l' = [l'_1, \ldots, l'_K] \) is given by:

\[
Q(l, l') = \begin{cases} 
\lambda_i, & \text{if } l_i + 1 = l'_i \text{ and } l_j = l'_j, \ j \neq i \\
 r_i(A_l), & \text{if } l_i - 1 = l'_i \text{ and } l_j = l'_j, \ j \neq i \\
- \sum_{l'' \neq l} Q(l, l''), & \text{if } l = l' \\
0, & \text{otherwise}
\end{cases}
\]

where the service rate \( r_i(A_l) \) is defined in Section III. The CTMC for a 2-BTS system is illustrated in Fig. 2.

As aforementioned, the steady-state distribution of the CTMC is unknown in general. In order to make progress, we shall approximate the original CTMC by a modified CTMC with reduced memory. Because the service rate \( r_i \) depends only on the set of active queues, we group all the states corresponding to any active set \( A \) and refer to them as group \( A \). The active set \( A_l \) can be thought of as a coarse binary quantization of state \( l \). According to (6), transitions occur between neighboring states where the length of a single queue increase or decrease by 1. Moreover, the CTMC makes a certain number of transitions between states within a group \( A \) before it jumps into a different group \( A' \). Specifically, we make the following two (somewhat strong) assumptions about the modified CTMC:

1) When the modified CTMC transits from some state \( l \) in group \( A \) to some state \( l' \) in a
different group \( A' \), the new state within \( A' \) is independent of \( l \), namely, such a transition is memoryless;

2) When the modified CTMC transits from group \( A \) to a different group \( A' \), the probability of assuming any state in \( A' \) is proportional to the stationary distribution of the state in the modified CTMC.

In the following, we analyze the stationary distribution of the modified CTMC. By definition, each inter-group state transition of the CTMC is a renewal, because the state is chosen anew within the new group according to the steady-state distribution. In addition, the intra-group transitions within a group form an independent CTMC. An important observation is that, within a group, where the set of active BTS’s is fixed, all service rates are invariant with the queue lengths, so that the \( K \) queues become independent. The probability of a state \( l \) is thus decomposed as

\[
p(l) = p(A_l) \prod_{i=1}^{K} p_i(l_i|A_l)
\]

where \( p_i(l_i|A) \) is the probability that queue \( i \) has length \( l \) given that the state is in group \( A \). Evidently, \( p_i(0|A) = 0 \) if \( i \in A \), \( p_i(l|A) = 0 \) if \( i \notin A, l > 0 \) and \( p_i(0|A) = 1 \) if \( i \notin A \). It suffices to determine the steady-state distribution of the groups \( A \in 2^S \), and, for each group \( A \), the steady-state distribution of the states in \( A \).

1) The Intra-group CTMC: Consider the independent CTMC within any given group \( A \). The CTMC can only exit group \( A \) when some queue length is equal to 1. For queue \( i \in A \), the probabilities of states 1, 2, \ldots must satisfy the detailed balance equation:

\[
p_i(l|A)\lambda_i = p_i(l+1|A)r_i(A), \quad l = 1, 2, \ldots
\]

As a result, for every \( i \in A \),

\[
p_i(l|A) = \left(1 - \frac{\lambda_i}{r_i(A)}\right)\left(\frac{\lambda_i}{r_i(A)}\right)^{l-1}, \quad l = 1, 2, \ldots
\]

It is easy to check that

\[
\sum_{l=1}^{\infty} p_i(l|A) = 1.
\]
2) The Inter-group CTMC: The inter-group transitions can be modeled by an inter-group CTMC with $2^K$ states, each corresponding to one group of active BTS’s $A$, and hence is referred to as a lumped state in view of the original CTMC. Two types of transitions can occur between the lumped states. First, the inter-group CTMC may transit from a lumped state $B$ where queue $i$ is empty to another lumped state $A$ where queue $i$ becomes nonempty, i.e.,

$$A = B \cup \{i\} \quad \text{and} \quad i \notin B.$$  \hspace{1cm} (11)

The rate of such a transition is the rate that a packet arrives at queue $i$, $\lambda_i$. Second, the inter-group CTMC may transit from a lumped state $A$ where queue $i$ has length 1 to another lumped state $B$ where queue $i$ becomes empty ($A$ and $B$ satisfy (11)). The rate of such a transition is the probability that the queue length is equal to 1 multiply the service rate, and can be expressed as:

$$p_i(1|A)r_i(A) = \left(1 - \frac{\lambda_i}{r_i(A)}\right)r_i(A)$$  \hspace{1cm} (12)

$$= r_i(A) - \lambda_i.$$  \hspace{1cm} (13)

The transition rates that completely describe the inter-group CTMC of the lumped states are expressed as:

$$\hat{Q}(A, B) = \begin{cases} 
\lambda_i, & \text{if } B = A \cup \{i\}, i \notin A \\
r_i(A) - \lambda_i, & \text{if } A = B \cup \{i\}, i \notin B \\
- \sum_{C: C \neq A} \hat{Q}(A, C), & \text{if } A = B \\
0, & \text{otherwise.}
\end{cases}$$  \hspace{1cm} (14)

The steady-state distribution $p(A)$ of the lumped states can be readily computed based on (14) using standard techniques in [17]. The key step therein is to invert a modified transition rate matrix.

As an example, the CTMC of the lumped states is illustrated for the 2-BTS case in Fig. 3.
3) **Average Delay:** The average length of queue $i$ in the modified CTMC can be calculated as follows:

$$\bar{l}_i = \sum_l l_i p(l)$$  \hspace{1cm} (15)

$$= \sum_l l_i p(A_l) \prod_{j=1}^K p_j(l_j|A_l)$$  \hspace{1cm} (16)

$$= \sum_A p(A) \sum_{l: A_l = A} l_i \prod_{j=1}^K p_j(l_j|A)$$  \hspace{1cm} (17)

$$= \sum_A p(A) \sum_{l_i: i \in A} l_i p_i(l_i|A) \left( \prod_{j \in A \setminus \{i\}} \sum_{l_j = 1}^\infty p_j(l_j|A) \right)$$  \hspace{1cm} (18)

$$= \sum_A p(A) \sum_{l_i: i \in A} l_i p_i(l_i|A)$$  \hspace{1cm} (19)

$$= \sum_{A: i \in A} p(A) \sum_{l=1}^\infty l p_i(l|A)$$  \hspace{1cm} (20)

$$= \sum_{A: i \in A} p(A) \frac{r_i(A)}{r_i(A) - \lambda_i}.$$  \hspace{1cm} (21)

By Little’s law, the average delay is given by $t_i = \bar{l}_i / \lambda_i$:

$$t_i = \sum_{A: i \in A} \frac{p(A) r_i(A)}{(r_i(A) - \lambda_i) \lambda_i}.$$  \hspace{1cm} (22)

The steady-state distribution of the modified CTMC [7] is not the same as that of the original interactive queueing system. Hence the average delay of the modified CTMC is in general an
approximation of that of the original CTMC. The accuracy of this refined approximation will be discussed in Section V-C.

B. The Optimization Problem

Using the approximate average delay given by (22), we formulate the refined $K$-BTS spectrum allocation problem as:

\[
\text{minimize}_{\{x(B), B \subseteq S\}} \sum_{i=1}^{K} \frac{\lambda_i}{\sum_{j=1}^{K} \lambda_j} t_i \tag{P2a}
\]

subject to

\[
t_i = \sum_{A : i \in A} \frac{p(A)r_i(A)}{(r_i(A) - \lambda_i)\lambda_i}, \quad \forall i \in S \tag{P2b}
\]

\[
r_i(A) = \sum_{B \subseteq S} s_i(B \cap A)x(B), \quad \forall i \in S, \quad \forall A \in 2^S \tag{P2c}
\]

\[
r_i(S) > \lambda_i, \quad \forall i \in S \tag{P2d}
\]

\[
x(B) \geq 0, \quad \forall B \in 2^S \tag{P2e}
\]

\[
\sum_{B \subseteq S} x(B) = 1. \tag{P2f}
\]

Constraint (P2d) gives the feasible region of the approximation due to the approximate steady state distribution in lumped state $S$. Since $P(A)$ is a function of $r_i(A)$’s, whose calculation involves a matrix inversion, it is difficult to check the convexity of the objective function (P2a). Simulations in Section VI used a standard convex optimization algorithm to solve the refined spectrum allocation problem, and appeared to converge to the same solution regardless of the initial point. Unlike Theorem 1 for the conservative spectrum allocation problem, the optimal solution to the refined spectrum allocation problem may divide the spectrum into more than $K$ segments.

C. Existing Bounds

The conservative approximation (3) is exactly the intuitive first-degree upper bound given in [9]. The corresponding first-degree lower bound can be derived by assuming all queues serve at the highest (single-user) rate, i.e., $r_i(\{i\})$. The first-degree upper (lower) bound is usually loose in the light (heavy) traffic regime since it ignores the interactions among queues.

\footnote{Here, the optimal solution means the locally optimal solution obtained by optimization solver in the numerical results.}
Tighter “second-degree” bounds are also presented in [9]. In the derivation of the second-degree bounds at BTS $i$, the fixed rate assumption is used to determine the utilization at other BTS’s, i.e., the fraction of the time that each BTS is transmitting. Then, the probability of each state of the rest of the network is calculated as the product of the corresponding busy/empty probabilities of the other BTS’s. To be specific, when deriving the second-degree upper bound for BTS $i$, all the other BTS’s are assumed to transmit at rate $r_j(S), j \neq i$. Denote $A_i(t)$ and $\bar{A}_i(t)$ as the set of active BTS’s exclusive of BTS $i$ at time $t$ in the original interactive queuing system and under this assumption, respectively. It can be proved using a sample path argument that $A_i(t) \subseteq \bar{A}_i(t)$ for every time instance $t$. Hence the service rate at BTS $i$ under this assumption is always lower than the service rate in the original interactive queuing system due to increased interference. Thus the worst-case rate assumption at the other BTS’s provides an upper bound. Similarly, in the derivation of the second-degree lower bound for BTS $i$, all other BTS’s are assumed to transmit at the highest rates, $r_j(\{j\}), j \neq i$.

When assuming fixed rates at other BTS’s, the queue at BTS $i$ reduces to a queue with time varying capacity. Processor sharing queues with time varying capacity was studied in [10], which showed that the upper and lower bounds on residual work load in such queues can be obtained by considering the quasi-stationary regime and the fluid regime, respectively. In the quasi-stationary regime, the rest of the system evolves so slowly that BTS $i$ only sees the initial states of the other BTS’s. On the other hand, in the fluid regime the rest of the system evolves so quickly that BTS $i$ only sees the average interference.

Under the worse-case transmit rate assumption, the probability that BTS $j, j \neq i$ transmits is:

$$\bar{p}_j = \frac{\lambda_j}{r_j(S)}.$$  \hspace{1cm} (23)

The probability that the other $K - 1$ BTS’s are in state $\bar{A}_i$ except BTS $i$ is:

$$\bar{\pi}_i(\bar{A}_i) = \prod_{j \in A_i} \bar{p}_j \prod_{l \in A_i, l \neq i} (1 - \bar{p}_l),$$  \hspace{1cm} (24)

The second-degree upper bound is finally given by taking the expectation over the distribution of all possible states $\bar{A}_i$:

$$\bar{t}_i = \sum_{\bar{A}_i \subseteq (S/\{i\})} \bar{\pi}_i(\bar{A}_i) \frac{1}{r_i(\bar{A}_i \cup \{i\}) - \lambda_i}.$$  \hspace{1cm} (25)
Under the highest transmit rate assumption, the probability of BTS $j$, $j \neq i$ being active is:

$$p_j = \frac{\lambda_j}{r_j(\{j\})}.$$  \hfill (26)

The corresponding probability of state $A_i$ for the $K - 1$ BTS’s except BTS $i$ is:

$$\pi_i(A_i) = \prod_{j \in A, j \neq i} p_j \prod_{l \notin A, l \neq i} (1 - p_j).$$ \hfill (27)

The second-degree lower bound is calculated using the average rate:

$$t_i = \frac{1}{\sum_{\Delta_i \in (S/\{i\})} \prod_{A_i} \pi_i(A_i) r_i(A_i \cup \{i\}) - \lambda_i}.$$ \hfill (28)

D. Average Delay and Stability Region

1) Average Delay: The accuracy of the refined approximation for the average delay developed in Section V-A is guaranteed by the following theorem:

**Theorem 2:** In a $K$-BTS interactive queueing system, the refined approximate mean packet sojourn time provided by (22) is between the second-degree upper and lower bounds in (25) and (28), i.e., $t_i < t_i < \bar{t}_i$.

The proof of *Theorem 2* is given in Appendix A.

2) Stability: The throughput region of the slow timescale spectrum allocation strategy is the same as the throughput region achieved by static inter-cell scheduling, which has been studied in [21]. The throughput region for the slow timescale spectrum allocation is:

$$\Lambda = \left\{(\lambda_1, \ldots, \lambda_K) \left| \exists x(B), \ B \in 2^S, \ \text{s.t.} \ r_i(S) \geq \lambda_i, \ \forall i \in S \right\} \right..$$  \hfill (29)

For any rate tuple in the interior of $\Lambda$, there exists a spectrum allocation that stabilizes the system, whereas for any rate tuple outside $\Lambda$, there is no spectrum allocation that stabilizes the system. Since the constraint $r_i(S) > \lambda_i$ $\forall i \in S$ is included in both (P1) and (P2), both the conservative and refined spectrum allocations achieve the maximum throughput region of the network.

The stability of $K$ interactive queues, with arbitrary Poisson arrivals and exponential service times is more complicated. For example, for a given $(\lambda_1, \ldots, \lambda_K)$ inside the throughput region (29), there may be a spectrum allocation $(x(B))_{B \in 2^S}$ other than the one given in (29), which also stabilizes the system but has $r_i(S) < \lambda_i$ at some queue $i$. The intuition is that although the worst-case rate at queue $i$ is less than the packet arrival rate, BTS $i$ could transmit at higher rate when some other BTS’s are inactive. As long as the average service rate exceeds the packet
arrival rate at each queue, the system will be stable. A sufficient condition is \( r_i(S) > \lambda_i \forall i \in S \).

This condition also characterizes the feasible region of the proposed conservative and refined approximations. A tighter sufficient condition has been shown in [9] using Loynes’ theorem [22]. The stability condition in heavy traffic limit for two interactive queues has been studied in [23].

**VI. Numerical Results**

**A. One Dimensional Example**

We first present a one dimensional example in Fig. 4 to illustrate that the proposed spectrum allocation scheme is topology and traffic aware. In this example, seven BTS’s are randomly dropped on a line segment represented by the blue triangles in Fig. 4. We refer to these BTS’s as BTS 1 to 7 from the left to the right. The parameters used in this simulation are: the path-loss exponent is 3, the transmit PSD is 1 watt/Hz for all BTS’s and the noise PSD is 0.125 micro-watt/Hz. This set of parameters were used for all the simulations in Section VI. The relative traffic load at each BTS is shown by the height of each red rectangle below it. The green and white rectangles above each BTS are the corresponding refined spectrum allocation. The rectangles colored green mean they are used by the corresponding BTS. Both spatial reuse and local orthogonalization are achieved by the refined allocation. The spectrum allocations at BTS 6 and 7 show that the refined allocation is traffic aware, since it assigns a larger portion of the spectrum to BTS 7, which has the heavier load of the two.
B. Two Dimensional Simulation Model

To illustrate the performance of the proposed conservative and refined spectrum allocations, we adopt the quantized HetNet model in [8]. A $100 \times 100$ m$^2$ area is quantized by hexagons with the distance between the centers of adjacent hexagons being 20 m. In the simulation, 7 BTS’s are uniformly randomly dropped at the vertices of the hexagons. UE locations within each hexagon are approximated by the center of the hexagon. UEs are assigned to their respective nearest BTS’s. We only consider path-loss in these simulations, although slow fading can be easily incorporated. Spectral efficiencies are calculated using Shannon’s capacity formula. The average spectral efficiency of BTS $i$ is calculated as the mean of the spectral efficiencies in the hexagons served by BTS $i$.

C. Delay Comparison

The refined and conservative approximations are compared with the second-degree upper and lower bounds in Fig. 5. The average delay versus average traffic arrival rate curves in Fig. 5 are based on the same spectrum allocation. The conservative approximation, also known as the first-degree upper bound, is coarser than the second-degree upper bound. As predicted by Theorem 2, the refined approximation is between the second-degree upper and lower bounds. In addition, the refined approximation is also quite accurate within the feasible region.

We compare the conservative and refined spectrum allocations with simpler orthogonal and
full-spectrum-reuse allocations for different traffic loads in Fig. 6. The orthogonal allocation is the optimal one among all feasible orthogonal allocations. Since the BTS’s do not interfere with each other under the orthogonal allocation, the average packet sojourn time is computed using the delay formula for the M/M/1 queue in (P1a). The delay performance for the other three allocations are achieved by simulation using the uniformization method [24]. The figure shows that the orthogonal allocation becomes unstable after the average packet arrival rate reaches 27 packets/second. The delay under full-spectrum-reuse allocation grows much faster after the average packet arrival rate rises above 21 packets/second. This turning point is related to the receive SNR, path-loss model and network topology. A general observation is that as the difference in spectral efficiency under the orthogonal and full-spectrum-reuse allocations grows larger, the turning point moves to the left towards lighter traffic. As suggested by the numerical results presented later, the full-spectrum-reuse allocation also becomes unstable at 27 pacekts/second. Hence, the proposed conservative and refined allocations achieve a larger throughput region than the other schemes. We can also observe from Fig. 6 that the proposed conservative and refined allocations achieve significant gains in the heavy traffic regime.

Since the delay under the full-spectrum-reuse allocation becomes large in the heavy traf-
fic regime, the orthogonal, conservative and refined allocations are compared separately in Fig. 7. Since the average sojourn time under the orthogonal allocation can be exactly calculated from (P1a), the minimum of (PI) given by the optimal conservative allocation is always no greater than the actual delay of the orthogonal allocation. Moreover, since the conservative approximation is an upper bound of the actual delay, the orthogonal allocation is always worse than the conservative allocation as shown in Fig. 7. The refined allocation always outperforms the conservative allocation due to the more accurate approximation of the actual delay. In the light traffic regime, the refined allocation reduces the average delay by about 60% compared to the conservative allocation; while both provide significant delay reduction compared to the orthogonal allocation. The advantage of using the refined allocation over the conservation allocation decreases as the traffic increases. In the heavy traffic regime, the difference is negligible.

The optimal conservative and refined allocations are shown in Fig. 8 for different traffic loads. The widths of the rectangles represent the values of nonzero $\chi(B)$’s. If the rectangle is red, it means that segment of the spectrum is used by the corresponding BTS of the row. By counting the number of nonzero segments on the left column, it is clear that the conservative allocation satisfies Theorem 1. In the light traffic regime, the refined allocation is close to full-spectrum-reuse, and as the traffic increases it becomes closer to the conservative allocation. This is because in the light traffic regime all $K$ BTS’s have very low utilization ratios, thus are inactive most of the time. In the heavy traffic regime, the $K$ interactive queues often have packets, and behave
average packet arrival rate = 3
average packet arrival rate = 9
average packet arrival rate = 15
average packet arrival rate = 21
average packet arrival rate = 27
average packet arrival rate = 33

Fig. 8. Optimal conservative (left) and refined (right) allocations with different packet arrival rates.

Fig. 9. BTS utilization ratios for different average packet arrival rates.

more like $K$ independent M/M/1 queues with the worst-case rates.

D. Utilization

We present the utilization ratio of each BTS under different traffic loads in Fig. 9. In Fig. 9a, where the traffic is very light, the refined allocation has very similar utilization as full-spectrum-reuse. The utilization ratios are much higher under the conservative allocation due to the worst-case (smallest) transmit rate assumption. In fact, for all four types of allocations the BTS’s are active less than 10% of the time, which also explains why orthogonal allocation is inefficient
in this light traffic regime. Fig. 9b presents the utilization ratios right after the turning point of the full-spectrum-reuse allocation as previously shown in Fig. 6. At 24 packets/second, the conservative and refined allocations have much lower BTS utilization ratios compared to the full-spectrum-reuse allocation, especially at BTS 3 and 5. Although the utilization ratios under orthogonal allocation are higher than that under the full-spectrum-reuse allocation, the orthogonal allocation actually has better delay performance at this traffic intensity. This is because all BTS’s always operating at the fixed rates under the orthogonal allocation, while the transmit rates vary according to the states of the system under the full-spectrum-reuse allocation. As the traffic demand increases, as shown in Fig. 9c, BTS’s 3 and 5 become saturated under full-spectrum-reuse, which also means the system becomes unstable. The conservative and refined allocations still maintain the stability of the queues and have much lower utilization ratios than the orthogonal allocation. As suggested in Fig. 6, if the traffic increases further to 30 packets/second, the orthogonal allocation also becomes unstable.

E. Delay Distribution

To evaluate the user QoS for different spectrum allocations, we show the CDF’s of number of packets in all queues for the entire network in Fig. 10. The 3 packets/second case in Fig. 10a represents the light traffic regime. The refined and full-spectrum-reuse allocations have the smallest number of packets at all percentiles. The conservative allocation has more packets than those two, but fewer packets than the orthogonal allocation. The figure also shows that 90% of the time the queues are empty under all four allocations. At around the turning point for full-spectrum-reuse in Fig. 6, the refined and conservative allocations both have much fewer packets.
at all percentiles than the other two, as shown in Fig. [10b]. When comparing the orthogonal and full-spectrum-reuse allocations, the orthogonal allocation has fewer time instances with a small number of packets and also fewer time instances with a lot of packets, i.e., a more balanced queue length. This is because the transmit rate at each BTS changes dramatically under the full-spectrum-reuse allocation according to the states of the system, whereas they stay the same under the orthogonal allocation. At even higher traffic loads, as shown in Fig. [10c], the full-spectrum-reuse allocation becomes unstable. The reason that the number of packets shown in the queue is finite is because we only simulated for $10^5$ time instances. The conservative and refined allocations have prominent advantages over the other two in this heavy traffic regime.

F. Power Control

The discussions so far have been based on the important assumption that the spectral efficiencies when shared by different combinations of BTS’s, i.e., $s_i(A), \forall i \in S, A \in 2^S$, are fixed. This assumption enables us to simplify the relationship between spectrum allocation and flow level service rate. In fact, the service rate is a linear function of the $x(B)$’s under this fixed spectral efficiency assumption. This assumption is valid if all BTS’s transmit with fixed power spectral density. However, in practice we may have a fixed total transmit power constraint at each BTS. Hence power control is not present in our formulation. For example, the spectral efficiencies for the orthogonal allocation should be higher than shown, since each BTS would concentrate all the transmit power on its exclusive spectrum.

The joint power control and spectrum allocation problem is in general more complicated. Hence, we take a simplified approach by alternatively updating the spectrum and power allocations. At the beginning, the $s_i(A)$’s are fixed assuming each BTS uniformly allocates its maximum transmit power across the entire spectrum. Then, we iterate the following steps:

1) Update the spectrum allocation $x(B), \forall B \in 2^S$ with the current $s_i(A), \forall i \in S, A \in 2^S$ by solving the proposed spectrum allocation problems.

2) Update the spectral efficiencies $s_i(A), \forall i \in S, A \in 2^S$ with the current $x(B), \forall B \in 2^S$ by letting each BTS $i$ uniformly allocate its maximum transmit power over the spectrum segments assigned to it, which include $x(B)$’s with $i \in B$.

The iterations continue until the $x(B)$’s converge (This is not guaranteed). The average packet sojourn time after the spectrum allocation update and the spectral efficiency update at each
iteration is shown in Fig. 11 for an average packet arrival rate per BTS of 24 packets/second. The figure shows that the delay performance converges very quickly for both conservative and refined allocations. The mean sojourn times for both allocations decrease substantially after the first spectral efficiency update. This is because at this average packet arrival rate both allocations orthogonalize the spectrum use among neighboring BTS’s to some extent. (The small variations in these curves are due to the limited simulation time.) This kind of convergence behavior can be expected in general, except for networks with very few BTS’s. This is because spatial reuse will occur in the conservative and refined allocations for relative large networks even in the heavy traffic regime. Since each BTS will use a fairly large amount of the spectrum, the spectral efficiencies will not change much after several iterations.

VII. CONCLUSIONS

Traffic aware spectrum allocation performed over a slow timescale for delay minimization has been studied in HetNets, where interference is caused only by active BTS’s. The model corresponds to interactive queues with coupled processors. Optimization problems were formulated using conservative and refined approximations of the mean packet sojourn time. The proposed model and queueing analysis for densely deployed HetNets depart from the traditional single cell and regular hexagonal cellular network models. Also, we have conditional on the topology of the network, instead of attempting to average over random realization as in [25].
According to the numerical results, the conservative and refined allocations significantly reduce the average delay by exploiting the network topology and the different loads across BTS’s. The problem formulation and results do not rely on the specific choice of Shannon spectral efficiency, which could take other forms. For future work, we plan to study a finer classification of UE service rates within a cell, e.g., the UEs located at cell center and cell edge are modeled as different types of users with different spectral efficiencies. Another direction is to provide scalable allocation methods that can be applied to large networks.

**APPENDIX A: PROOF OF THEOREM 2**

First we will prove $t_i < \bar{t}_i$. Both (22) and (25) can be regarded as weighted sum of $\frac{r_i(A)}{(r_i(A) - \lambda_i)\lambda_i}$. The weight is the probability of the system being in state $A$ under different approximation. Based on the worst-case rate assumption for deriving $\bar{t}_i$, we can form a similar lumped chain as (14) with the transition rate matrix:

$$Q(A, B) = \begin{cases} 
\lambda_i, & \text{if } B = A \cup \{i\}, i \notin A \\
(r_i(S) - \lambda_i), & \text{if } A = B \cup \{i\}, i \notin B \\
- \sum_{C \neq A} Q(A, C), & \text{if } A = B \\
0, & \text{otherwise.}
\end{cases}$$

(30)

The lumped chains (14) and (30) can be interpreted as the CTMC of the following Markov process: There are $K$ interactive queues each with capacity 1. Packet arrivals at queue $i$ follow a Poisson process with rate $\lambda_i$. The size of each packet is independently exponentially distributed with unit mean. The service rates are given by the corresponding rate matrices of (14) and (30), which are state dependent. If a packet arrives at an empty queue, it is immediately served. Otherwise, the packet in service will be discarded and the newly arrived packet immediately starts being served. Denote $S$ and $\bar{S}$ respectively as the corresponding systems with probability transition matrix given by (14) and (30). Let $A(t)$ and $\bar{A}(t)$ be the sets of active queues at time $t$. Denote the residual loads (in bits) by $l(t)$ and $\bar{l}(t)$, where $l(t), \bar{l}(t) \in \mathbb{R}_+^K$ and the $i$th elements $l_i(t)$ and $\bar{l}_i(t)$ are the residual loads in queue $i$ of the two systems. Using a sample path argument, we can show:

$$A(t) \subseteq \bar{A}(t)$$

$$l(t) \leq \bar{l}(t)$$

(31)
at any time instance \( t \). The inequality \( \mathbf{l}(t) \leq \bar{\mathbf{l}}(t) \) is element-wise, i.e., \( l_i(t) \leq \bar{l}_i(t) \), \( i = 1, \ldots, K \).

Assuming both systems evolve under the same packet arrival realization, we can prove (31) by induction. At \( t = 0 \), \( A(0) = \bar{A}(0) = \emptyset \) and \( \mathbf{l}(0) = \bar{\mathbf{l}}(0) = 0 \). Assume (31) is true at \( t = \tau > 0 \). Then, (31) still holds before any arrival or departure happens at \( t = \tau + \delta \). This is because \( A(t) \subseteq \bar{A}(t) \), \( \forall t \in [\tau, \tau + \delta) \) implies the service rate of any queue \( i \) in \( S \) is larger than the service rate of queue \( i \) in \( \bar{S} \) within this time interval. At \( t = \tau + \delta \), we have two cases.

Case 1. An arrival happens at queue \( j \). Since both systems follows the same sample path of the arrivals, we still have \( A(\tau + \delta) \subseteq A(\tau + \delta) \). The residual load at other queues does not change from \( \tau + \delta_\_ \) to \( \tau + \delta \). At queue \( j \), \( l_j(\tau + \delta) = \bar{l}_j(\tau + \delta) \).

Case 2. A departure happens at queue \( j \). Since \( l(t) \leq \bar{l}(t) \), \( \forall t \in [\tau, \tau + \delta) \) and the service rate at each queue in \( S \) is no less than the service rate at the same queue in \( \bar{S} \), this departure must be in \( S \) (It could be the case that a departure also happens at queue \( j \) in \( \bar{S} \) at \( t = \tau + \delta \)). It immediately implies (31) at \( t = \tau + \delta \).

Because the Markov processes in both systems are ergodic, we have:

\[
\begin{align*}
t_i &= \lim_{T \to \infty} \frac{1}{T} \int_0^T f_i(A(t)) \, dA(t) \\
\tilde{t}_i &= \lim_{T \to \infty} \frac{1}{T} \int_0^T f_i(\bar{A}(t)) \, d\bar{A}(t),
\end{align*}
\]

where the function \( f_i(A) \) is:

\[
f_i(A) = \begin{cases} 
\frac{r_i(A)}{(r_i(A) - \lambda_i)\lambda_i} & \text{if } i \in A \\
0 & \text{otherwise}
\end{cases}
\]

It is easy to see that if \( A \subseteq B \) then \( f_i(A) \leq f_i(B) \). Applying this and (31) to (32), we can obtain \( t_i < \tilde{t}_i \), where the strict inequality is due to ergodicity.

To prove \( t_i > \tilde{t}_i \), we first introduce an intermediate variable \( t'_i \):

\[
t'_i = \frac{1}{\sum_{A : i \in A} P(A) \frac{r_i(A)}{\lambda_i} r_i(A) - \lambda_i}.
\]

According to (14), it is easy to check \( \sum_{A : i \in A} P(A) \frac{r_i(A)}{\lambda_i} = 1 \). By Jensen’s inequality, we immediately have \( t_i \geq t'_i \). Hence, we only need to show \( \tilde{t}_i > t'_i \). Again we can form a lumped
chain with the highest service rate assumption, whose transition rate matrix is:

\[
Q(A, B) = \begin{cases} 
\lambda_i, & \text{if } B = A \cup \{i\}, i \notin A \\
(r_i(\{i\}) - \lambda_i), & \text{if } A = B \cup \{i\}, i \notin B \\
-\sum_{C:C \neq A} Q(A, C), & \text{if } A = B \\
0, & \text{otherwise.}
\end{cases}
\]  

(34)

Using a similar sample path argument we can prove the expected service rate in (33) is less than the expected service rate in (28), which directly implies \( t'_i > t_i \).

REFERENCES


