Statistical Multiplexing and Traffic Shaping Games for Network Slicing

Jiaxiao Zheng*, Pablo Caballero*, Gustavo de Veciana*, Seung Jun Baek† and Albert Banchs‡

*The University of Texas at Austin, TX
†Korea University, Korea
‡University Carlos III of Madrid & IMDEA Networks Institute, Spain
contact: gustavo@ece.utexas.edu

Abstract—Next generation wireless architectures are expected to enable slices of shared wireless infrastructure which are customized to specific mobile operators/services. Given infrastructure costs and the stochastic nature of mobile services’ spatial loads, it is highly desirable to achieve efficient statistical multiplexing amongst network slices. We study a simple dynamic resource sharing policy which allocates a ‘share’ of a pool of (distributed) resources to each slice—Share Constrained Proportionally Fair (SCPF). We give a characterization of the achievable performance gains over static slicing, showing higher gains when a slice’s spatial load is more ‘imbalanced’ than, and/or ‘orthogonal’ to, the aggregate network load. Under SCPF, traditional network dimensioning translates to a coupled share dimensioning problem, addressing the existence of a feasible share allocation given slices’ expected loads and performance requirements. We provide a solution to robust share dimensioning for SCPF-based network slicing. Slices may wish to unilaterally manage their users’ performance via admission control which maximizes their carried loads subject to performance requirements. We show this can be modeled as a “traffic shaping” game with an achievable Nash equilibrium. Under high loads the equilibrium is explicitly characterized, as are the gains in the carried load under SCPF vs. static slicing. Detailed simulations of a wireless infrastructure supporting multiple slices with heterogeneous mobile loads show the fidelity of our models and range of validity of our high load equilibrium analysis.

I. INTRODUCTION

Next generation wireless systems are expected to embrace SDN/NFV technologies towards realizing slices of shared wireless infrastructure which are customized for specific mobile services e.g., mobile broadband, media, OTT service providers, and machine-type communications. Customization of network slices may include allocation of (virtualized) resources (communication/computation), per-slice policies, performance monitoring and management, security, accounting, etc. The ability to deploy service specific slices is viewed, not only as means to meet the diverse and sometimes stringent demands of emerging services, e.g., vehicular, augmented reality, but also an approach for infrastructure providers to reduce costs while developing revenue streams. Resource allocation in this context is more challenging than for traditional cloud computing. Indeed, rather than drawing on a centralized pool of resources, a network slice requires allocations across a distributed pool of resources, e.g., base stations. The challenge is thus to promote efficient statistical multiplexing amongst slices over pools of shared resources.

The focus of this paper will be resource sharing amongst slices supporting stochastic (mobile) loads. A natural approach to sharing is static slicing, whereby resources are statically partitioned and allocated to slices. This offers each slice a guaranteed allocation at each base station, and protection from each other’s traffic, but, as we will see, poor efficiency. Instead, we consider, an alternative wherein each slice is pre-assigned a fixed share of the pool of resources, and redistributes its share equally amongst its active customers. In turn, each base station allocates resources to customers in proportion to their shares. We refer to this sharing model as Share Constrained Proportionally Fair (SCPF) resource allocation. By contrast with static slicing, SCPF is dynamic (since its resource allocations depend on the network state) but constrained by the network slices’ pre-assigned shares (which provides a degree of protection amongst slices).

Related work. There is an enormous amount of related work on network resource sharing in the engineering, computer science and economics communities. The standard framework used in the design and analysis of communication networks is utility maximization (see e.g., [20] and references therein) which has led to the design of several transport and scheduling mechanisms and criteria, e.g., the often considered proportional fair criterion. The SCPF mechanism, described above, should be viewed as a Fisher market where agents (slices), which are share (budget) constrained, bid on network resources, see, e.g., [17] and for applications [3], [11]. The choice to re-distribute a slice’s shares (budget) equally amongst its users, can be viewed as a network mandated policy, but also emerges naturally as the social optimal, market and Nash equilibrium when slices exhibit (price taking) strategic behavior in optimizing their own utility, see [7].

The novelty of our work lies in considering slice based sharing, under stochastic loads and in particular studying the expected performance resulting from such SCPF-based coupling slices’ customer allocations. Other researchers who have considered performance of stochastic networks, e.g., [5], [9] and others, have studied networks where customers are allocated resources (along routes) based on maximizing a sum of customers utilities. These works focus on network stability for ‘elastic’ customers, e.g., file transfers. Subsequently [6], [19] extended this line of work, to the evaluation of mean file delays, but only under balanced fair resource allocations.
Each slice $v$ customer at base station $b$ has an independent sojourn time with mean $\mu^v_b$ after which it is randomly routed to another base station or exits the system. As explained below we assume that such mobility patterns do not depend on the resources allocated to users. We let $Q^v = (q^v_{i,j} : i, j \in B)$ denote a slice-dependent routing matrix where $q^v_{i,j}$ is the probability a slice $v$ customer moves from base station $i$ to $j$ and $1 - \sum_{j \in B} q^v_{i,j}$ is the probability it exits the system. Throughout the paper, we assume $Q^v$ is irreducible for all $v \in V$. This model induces an overall traffic intensity for slice $v$ across base stations satisfying flow conservation equations: for all $b \in B$ we have

$$\kappa^v_b = \gamma^v_b + \sum_{a \in B} \kappa^v_a q^v_{a,b},$$

where $\kappa^v_b$ is the traffic intensity of slice $v$ on base station $b$. Accounting for users’ sojourn times, the mean offered load of slice $v$ on base station $b$ is $\rho^v_b = \kappa^v_b \mu^v_b$, and $\rho^v = (\rho^v_b : b \in B)^T$ captures its overall system load. Letting $\mu^v = (\mu^v_b : b \in B)^T$, the flow conservation equations can be rewritten in matrix form as:

$$\rho^v = \text{diag}(\mu^v)(I - (Q^v)^T)^{-1}\gamma^v.$$

Note that $I - (Q^v)^T$ is irreducibly diagonally dominant and thus invertible.

This model corresponds to a multi-class network of $M/GI/\infty$ queues (base stations), where each slice corresponds to a class of customers, see, e.g., [14]. Such networks are known to have a product-form stationary distribution, i.e., the numbers of customers on slice $v$ at base station $b$ denoted by $N^v_b$ are mutually independent and $N^v_b \sim \text{Poisson}(\rho^v_b)$. Since the sum of independent Poisson random variables is again Poisson, the total number of customers on slice $v$ is such that $N^v = \sum_{b \in B} N^v_b \sim \text{Poisson}(\rho_v)$ where $\rho_v = \sum_{b \in B} \rho^v_b$.

Our network model for the numbers of customers and mobility across base stations, assumes customer sojourn/activity/mobility are independent of the network state and of the resources a customer is allocated. This is reasonable for properly engineered slices where the performance a customer sees does not impact its activity, e.g., inelastic or rate adaptive applications seeing acceptable performance. This covers a wide range of applications including voice, video streaming, IoT monitoring, real-time control, and even web browsing sessions experiencing good performance. This model however is not appropriate for customers sensitive to file download delays, e.g., who might leave the system earlier if allocated more resources.

There are several natural generalizations to this model including class-based routing and user sessions (e.g. web browsing) which are not always active at the base stations they visit, see e.g., [14].

B. Network Slice Resource Sharing

In the sequel we consider a setting where the resources allocated to a slice’s customers depend on the overall network state, i.e., number of customers each slice has on each base station, corresponding to the stochastic process described in...
Section II-A. Let us consider a snapshot of the system’s state and let $U^b_v, U_b, U^v$ and $U$ denote sets of active customers on slice $v$ at base station $b$, at base station $b$, on slice $v$ and on the overall network respectively. Thus, the cardinalities of these sets correspond to a realization of the system ‘state’, i.e., $|U^b_v| = n^v_b$ and $|U^v| = n^v$, where in a stationary regime $n^v$ and $n^v_b$ are realizations of Poisson random variables $N^v$ and $N^v_b$, respectively.

Each base station $b$ is modeled as a finite resource shared by its associated users $U_b$. A customer $u \in U_b$ can be allocated a fraction $f_u \in [0, 1]$ of that resource, e.g., of resource blocks in a given LTE frame, or allocated the resource for a fraction of time, where $\sum_{u \in U_b} f_u = 1$. We shall neglect quantization effects. The transmission rate to customer $u$, denoted by $r_u$, is then given by $r_u = f_u c_u$ where $c_u$ denotes the current peak rate for that user. To model customer heterogeneity across slices/base stations we shall assume $c_u$ is a typical customer on slice $v$ at base station $b$ is an independent realization of a random variable with the same distribution as $C^v_b$. It may depend on the slice, since slices may support different types of customer devices (e.g., car connectivity vs mobile phone) and depend on the base station, since typical slice $v$ users may have different spatial distributions with respect to base station $b$ or see different levels of interference.

Below we consider two resource allocation schemes. For both we assume each slice shares a ‘share’ of the network resources $s_v, v \in V$ such that $s_v > 0$ and $\sum_{v \in V} s_v = 1$. Let $\tilde{\rho}_b$ denote the load demand of base station $b$, at base station $b$. Such averages naturally place a higher weight on congested base stations, where a slice may have more users, best reflecting the overall performance customers will see.

### Definition 1. Static Slicing (SS): Under SS, slice $v$ is allocated a fixed fraction $s_v$ of each base station $b$’s resources, and each customer $u \in U^v_b$ gets an equal share, i.e., $1/n^v_b$, of the slice $v$’s resources at base station $b$. Thus the users transmission rate $r^SS_u$ is given by

$$r^SS_u = \frac{s_v}{n^v_b} c_u.$$ 

### Definition 2. Share Constrained Proportionally Fair (SCPF): Under SCPF each slice re-distributes its share of the overall network resources equally amongst its active customers, which thus get a sub-share (weight) $w_u = \frac{s_v}{n^v_b}$ for $u \in U^v, \forall v \in V$. In turn, each base station allocates resources to customers in proportion to their weights. So a user $u \in U^v_b$ gets a transmission rate $r^SCPF_u$ given by

$$r^SCPF_u = \sum_{u' \in U^v_b} \frac{w_u}{n^v_b} c_u = \frac{s_v}{\sum_{v' \in U^v} \frac{n^v_{v'}}{n^v} s_{v'}} c_u.$$ 

Thus under SCPF the overall fraction of resources slice $v$ is allocated at a base station $b$ is proportional to $n^v_b/s_v$, i.e., its share and its relative number of users at the base station. This provides a degree of elasticity to variations in the slice’s spatial loads. However, if a slice has a large number of customers, its customers’ weights are proportionally decreased, which protects other slices. In addition to being quite simple to implement, as mentioned in Section I SCPF resource allocations are socially optimal, and correspond to market and Nash equilibria for certain types of budget-constrained Fisher Markets.

### III. PERFORMANCE EVALUATION

In this section we study the expected performance seen by a slice’s typical customer. Given our focus in inelastic/rate adaptive traffic and tractability, we choose our customer performance metric as the reciprocal transmission rate, referred to as the Bit Transmission Delay (BTD), see e.g., [21]. This corresponds to the time taken to transmit a ‘bit’, so lower BTDs indicate higher rates and thus better performance. Short packet transmission delays are roughly proportional to the BTD. Alternatively the negative of the BTD can be viewed as a concave utility function of the rate, which in the literature was referred as the potential delay utility. Given the stochastic loads on the network, we shall evaluate the average BTD seen by a typical (i.e., randomly selected) customer on a slice, i.e., averaged over the stationary distribution of the network state and transmission capacity seen by typical users, e.g., $C^v_b$, at each base station. Such averages naturally place a higher weight on congested base stations, where a slice may have more users, best reflecting the overall performance customers will see.

#### A. Analysis of BTD Performance

Consider a typical customer on slice $v$ and let $E^v$ denote the expectation of the system state as seen by such a customer, i.e., the Palm distribution [2]. For SCPF, we let $R^v$ be a random variable denoting the rate of a typical customer on slice $v$, and $R^v_b$ that of such customer of slice $v$ at base station $b$. Similarly, let $R^v,SS$ and $R^v,SS_b$ denote these quantities under static slicing. Thus, under SCPF the average BTD for a typical slice $v$ customer is given by $E^v[\frac{1}{R^v}]$. The next result characterizes the mean BTD under SCPF and SS under our traffic model.

We introduce some further notation: the normalized load distribution of slice $v$ is $\tilde{\rho}^v = (\tilde{\rho}^v_b : b \in B)^T$ where $\tilde{\rho}^v = \frac{\rho^v}{\rho_c}$; the overall share weighted normalized load distribution is $\tilde{\rho} = (\tilde{\rho}^v_b : b \in B)^T$ where $\tilde{\rho} = \sum_{v \in V} s_v \tilde{\rho}^v$; and the mean reciprocal resource capacity for slice $v$ is $\delta^v = (\delta^v_b : b \in B)^T$ where $\delta^v_b = E^v[\frac{1}{C^v_b}]$.

**Theorem 1.** For network slicing based on SCPF, the mean BTD for a typical customer on slice $v$ is given by

$$E^v \left[ \frac{1}{R^v} \right] = \sum_{b \in B} \tilde{\rho}^v_b \delta^v_b \left( 1 - \tilde{\rho}^v_b + \frac{(\rho_c + 1)}{s_v} \tilde{\rho}^v_b \right).$$

For network slicing based on SS, the mean BTD for a typical customer on slice $v$ is given by

$$E^v \left[ \frac{1}{R^{v,SS}} \right] = \sum_{b \in B} \tilde{\rho}^v_b \delta^v_b \left( \frac{\rho_c + 1}{s_v} \right).$$
as given by Eq. (2), the BTD of a typical slice $v$ user at base station $b$ can be expressed as follows:

$$\mathbb{E}^v \left[ \frac{1}{R^v_b} \right] = \mathbb{E}^v \left[ \frac{1}{C^v_b} \right] \mathbb{E} \left[ \frac{s_v^N \sum_{v' \neq v} s_v^N}{N^v + 1} \right] = \delta_b^v \mathbb{E} \left[ (N^v_b + 1) \frac{N + 1 + \sum_{v' \neq v} s_v^N N^v}{s_v} \right] = \delta_b^v \left( 1 - \rho^v_b + \frac{\rho^v + 1}{s_v} \tilde{g}_b \right).
$$

Where the last equality follows by noticing that (i) $N^v$ is independent of $N^v_b$ and $N^v'$ and (ii) $\mathbb{E} \left[ \frac{N^v}{N^v'} \right] = \frac{\rho^v_b}{\rho^v}$. The latter result is a generalization of the following observation using the infinite divisibility of Poisson random variables: suppose $X_1, X_2$ i.i.d. Poisson($\lambda$), then by symmetry we have

$$1 = \mathbb{E} \left[ \frac{X_1 + X_2}{X_1 + X_2} \right] = 2 \mathbb{E} \left[ \frac{X_1}{X_1 + X_2} \right] = \mathbb{E} \left[ \frac{X_1}{X_1 + X_2} \right] = \frac{1}{2}.
$$

Under static slicing we have that

$$\mathbb{E}^v \left[ \frac{1}{R^v_{b,SS}} \right] = \mathbb{E}^v \left[ \frac{1}{C^v} \right] \mathbb{E} \left[ \frac{N^v + 1}{s_v} \right] = \delta_b^v \rho^v_b + \frac{1}{s_v}.$$

The theorem follows by taking an averaged weight across base stations – weighted by the fraction of customers at each base station, i.e., $\delta_b^v$.

**IV. PERFORMANCE MANAGEMENT**

In practice each slice $v \in \mathcal{V}$ may wish to provide service guarantees to its customers, i.e., ensure that the mean BTD does not exceed a performance target $d_v$. Below we investigate how to dimension network shares to support slice loads subject to such mean BTD requirements.

**A. Share Dimensioning under SCPF**

Consider a network supporting the traffic loads of a single slice, say $v$, so $s_v = 1$ and $\tilde{g} = \tilde{\rho}^v$ and let $d_v \triangleq d_v/\delta_v$ denote slice $v$’s normalized BTD constraint. Note that $d_v$ is the minimum BTD achievable when a user gets all the base station resources, so a target requirement satisfies $d_v > d_v$, and so $d_v > 1$. For slice $v$ to meet a mean BTD constraint $d_v$, it follows from Eq. (3) that:

$$\rho_v \leq l(d_v/\tilde{\rho}^v) \triangleq d_v - \frac{1}{\|\tilde{\rho}^v\|_2^2}.
$$

We can interpret $(d_v/\tilde{\rho}^v)$ as the maximal admissible carried load $\rho_v$ given a fixed relative load distribution $\tilde{\rho}^v$ and requirement $d_v$. As might be expected, if the relative load distribution $\tilde{\rho}^v$ is more balanced, i.e., $\|\tilde{\rho}^v\|_2^2$ is smaller, or if the BTD constraint is relaxed, i.e., $d_v$ is higher, the slice can carry a higher overall load $\rho_v$.

Next, let us consider SCPF based sharing amongst a set of slices $\mathcal{V}$ each with its own BTD requirements. It follows from Eq. (3) that to meet such requirements on each slice the following should hold: for all $v \in \mathcal{V}$

$$s_v \geq \frac{1 + \rho_v}{l((d_v/\tilde{\rho}^v) - \rho_v)} \sum_{u \neq v} s_u \frac{\|\tilde{\rho}^u\|_2^2}{\|\tilde{\rho}^v\|_2^2} \cos(\theta(\tilde{\rho}^u, \tilde{\rho}^v)).
$$

(5)
This can be written as:

$$\sum_{v \in V} s_v \mathbf{h}^v \geq 0,$$

(6)

where we refer to $\mathbf{h}^v = (h^v_u : u \in V)^T$ as $v$’s share coupling vector, given by

$$h^v_u = \begin{cases} 1 & u = v \\ \frac{1 + \rho_u}{l(d_u, \tilde{\rho}^v) - \rho_u} \frac{\|\tilde{\rho}^u\|_2^2}{\|\tilde{\rho}^v\|_2^2} \cos(\theta(\tilde{\rho}^u, \tilde{\rho}^v)) & v \neq u. \end{cases}$$

We can interpret $h^v_u$ as the benefit to slice $v$ of allocating unit share to it. When $v \neq u$, $h^v_u$ depends on two factors. The first factor, $\frac{1 + \rho_u}{l(d_u, \tilde{\rho}^v) - \rho_u}$, captures the sensitivity of slice $u$ to the ‘share weighted congestion’ from other slices. If $\rho_u$ is close to its limit $l(d_u, \tilde{\rho}^v)$, its sensitivity is naturally very high. The second term, $\frac{\|\tilde{\rho}^u\|_2^2}{\|\tilde{\rho}^v\|_2^2} \cos(\theta(\tilde{\rho}^u, \tilde{\rho}^v))$ captures the impact of slice $v$’s load distribution on slice $u$. Note that if two slices load distributions are orthogonal, they do not affect each other.

The following result summarizes the above analysis.

**Theorem 2.** There exists a share allocation such that slice loads and BTD constraints $(\rho^v, \tilde{\rho}^v, d_v) : v \in V)$ are admissible under SCFP sharing if and only if there exists an $s = (s_v : v \in V)^T$ such that $\|s\|_1 = 1$, $s \succeq 0$ and

$$\sum_{v \in V} s_v h^v \geq 0.$$

Admissibility can then be verified by solving the following maximin problem:

$$\max_{s \succeq 0} \left\{ \min_{v \in V} \sum_{v \in V} s_v h^v : \|s\|_1 = 1 \right\}.$$  

(7)

If the optimal objective function is positive, the traffic pattern is admissible. Moreover, if there are multiple feasible share allocations, then the optimizer is a ‘robust’ choice in that it maximizes the minimum share given to any slice, giving slices margins to tolerate perturbations in the slice loads satisfying Eq. (6).

If a set of network slice loads and BTD constraints are not feasible, admission control will need to be applied. We discuss this in the next section.

**B. Admission Control and Traffic Shaping Games**

A natural approach to managing performance in overloaded systems is to perform admission control. In the context of slices supporting mobile services where spatial loads may vary substantially, this may be unavoidable. Below we consider admission control policies that adapt to changes in load. Specifically, an admission control policy for slice $v$ is parameterized by $\mathbf{a}^v = (a^v_b : b \in B)^T \in [0, 1]^B$ where $a^v_b$ is the probability a new customer at base station $b$ is admitted. Such decisions are assumed to be made independently thus admitted customers for slice $v$ at base station $b$ still follow a Poisson Process with rate $\gamma_b^v a^v_b$. Based on the flow conservation equation Eq. (1) one can obtain the carried load $\rho^v$ induced by admission control policy $\mathbf{a}^v$ via

$$\rho^v = (M^v)^{-1} \mathbf{a}^v = \text{diag}(\mu_v)(I - (Q^v)^T)^{-1} \text{diag}(\gamma_v) \mathbf{a}^v,$$

where $M^v = \text{diag}(\gamma_v)^{-1}(I - (Q^v)^T)\text{diag}(\mu_v)^{-1}$ is invertible because $I - (Q^v)^T$ is irreducibly diagonally dominant. By contrast with Section II-A, note that $\rho^v$ now represents the load after admission control, which may have a reduced overall load and possibly changed relative loads across base stations – i.e., shape the traffic on the slice. We also let $\tilde{\rho}$ be the overall share weighted relative loads after admission control, see Section III-A. Note that we have assumed only exogenous arrivals can be blocked, thus once a customer is admitted it will not be dropped – the intent is to manage performance to maintain service continuity.

Below we consider a setting where slices unilaterally optimize their admission control policies in response to network congestion, rather than a single joint global optimization. The intent is to allow slices (which may correspond to competing virtual operator/services) to optimize their own performance, and/or enable decentralization in settings with SCFP based sharing.

Suppose each slice $v$ optimizes its admission control policy so as to maximize its overall carried load $\rho_v$, i.e., the average number of active users on the network, subject to a mean BTD constraint $d_v$. Under Assumption 1 the optimal policy for slice $v$ is the solution to the following optimization problem:

$$\max_{\rho^v, \rho_v} \rho_v$$

s.t. $\mathbf{a}^v = \rho_v M^v \tilde{\rho}^v, \quad \rho^v \in [0, 1]^B, \quad 1^T \tilde{\rho}^v = 1$  

(9)

$$\frac{(\rho_v + 1)}{s_v} \tilde{g}^T \tilde{\rho}^v - \|\tilde{\rho}^v\|_2^2 \leq d_v - 1.$$  

(10)

Note that Eq. (9) establishes a one-to-one mapping between $(\tilde{\rho}^v, \rho_v)$ and $\mathbf{a}^v$. We will use $\tilde{\rho}^v$ and $\rho_v$ to parameterize admission control decisions for slice $v$. The BTD constraint in Eq. (10) follows from Eq. (3). Also note that this admission control policy depends on both the overall share weighted loads, if for all $v \in V$ we have $\rho_v \gg 1$.

Under Assumption 2 we have that $1 + \rho_v \approx \rho_v$, and the left hand side of Eq. (10) becomes:

$$\left(\frac{\rho_v + 1}{s_v}\right) \tilde{g}^T \tilde{\rho}^v - \|\tilde{\rho}^v\|_2^2 \approx \frac{\rho_v}{s_v} \tilde{g}^T \tilde{\rho}^v = (s_v x_v)^{-1} \tilde{g}^T \tilde{\rho}^v$$

(11)

where we have defined $x_v \triangleq \rho_v^{-1}$. Further defining $\tilde{\rho}^{-v} \triangleq (\tilde{\rho}^v : v' \in V \setminus \{v\})$, Eq. (10) can be replaced by:

$$f_v(\tilde{\rho}^{-v} ; \tilde{\rho}^v) \triangleq \tilde{g}^T \tilde{\rho}^v \leq s_v(d_v - 1)x_v.$$  

(12)

Thus, defining $y^v \triangleq (\tilde{\rho}^v, x_v)$, which is equivalent to $(\tilde{\rho}^v, \rho_v)$, together with $y^{-v} \triangleq (y^v : v' \in V \setminus \{v\})$, under Assumption 2

1If $\gamma_v$ is not strictly positive one can reduce the dimensionality.
each slice can unilaterally optimize its admission control policy by solving the following problem:

**Admission control for slice $v$ under SCPF($AC_v$):** Given other slices' admission decisions $y^*\setminus v$, slice $v$ determines its admission control policy $y^v = (\hat{\rho}^v, x^v)$ by solving

$$
\min_y \left\{ x_v \mid y^v \in Y^v(y^*\setminus v) \right\}
$$

where $Y^v(y^*\setminus v)$ denotes slice $v$'s feasible policies and is given by

$$
Y^v(y^*\setminus v) \triangleq \left\{ y^v \mid 1^T \hat{\rho}^v = 1, 0 \preceq M^v \hat{\rho}^v \preceq x_v 1, f_v((\hat{\rho}^v); \tilde{\rho}^v) \leq s_v(d_v - 1)x_v \right\}.
$$

Eq. (13) and (14) can be viewed as defining a game where each slice is a player, wishing to minimize a cost $x_v$, and constrained to a strategic space. Such a game has a Nash equilibrium if there exists a joint strategy $y^* = (y^{v*}, v \in V)$ such that no slice $v$ can unilaterally decrease its cost $x_v$. The following result follows from Theorem 3.1 in [10].

**Theorem 3.** The traffic shaping game defined above has a Nash equilibrium.

Next we study the characteristics of the resulting traffic shaping Nash equilibrium. To make this tractable we consider networks which are saturated and subsequently (Section V) provide simulations to evaluate other settings.

**Assumption 3.** (Saturated Regime) Suppose the system is such that for each network slice, the optimal admission control for both SCPF and SS in response to other slices' load is such that for all $v \in V$, $a^v \prec 1$.

Assumption 3 depends on many factors including the BTD constraint, the mobility pattern and network slices' share patterns, but it is generally true when the exogenous traffic of all slices at all base stations $\gamma^v_k$ is high. This is the case we have the following:

**Theorem 4.** Under Assumptions 1, 2 and 3, the relative load distributions at the Nash equilibrium of the traffic shaping game $\hat{\rho}^* \triangleq (\hat{\rho}^{v*}; v \in V)$ are the unique solution to:

$$
\min_{(\hat{\rho}^v \in \Gamma^v; v \in V)} \left\{ \sum_v s_v \hat{\rho}^v \parallel_2^2 + \sum_v s_v^2 \parallel \hat{\rho}^v \parallel_2^2 \right\},
$$

where $\Gamma^v \triangleq \left\{ \tilde{\rho}^v \mid 1^T \tilde{\rho}^v = 1, M^v \tilde{\rho}^v \succeq 0 \right\}$, and the associated carried load for slice $v$ is $\rho^v_* = s_v(d_v - 1)/\hat{g}^v_*, \tilde{\rho}^{v*}_*$ corresponds to the overall share weighted relative loads distributions at the equilibrium.

The proof of Theorem 6 follows directly by comparing the Karush-Kuhn-Tucker (KKT) conditions for Eq. (15) versus those associated with slices’ admission control problems. Furthermore, in the saturated regime, BTD constraints are binding so the total carried load can be obtained from Eq. (12). A detailed proof is included in the extended version of this paper, see [23].

The first term in the objective function in Eq. (15) rewards balancing the overall share weighted relative loads on network. The second term rewards a slice for balancing its own relative loads. The Nash equilibrium in the saturated regime is thus a compromise between these two objectives while constrained by the network slices mobility patterns and feasible admission control policies.

**Admission control for slice $v$ under SS ($AC_{v,SS}$):** Under SS slice $v$ can determine its optimal admission control $y^v$ by solving:

$$
\max_{\hat{\rho}^v, \rho_v} \rho_v
$$

s.t. $a^v = \rho_v M^v \tilde{\rho}^v$, $a^v \in [0, 1]^B$

$$
1^T \tilde{\rho}^v = 1 \quad \text{and} \quad \rho_v \parallel \tilde{\rho}^v \parallel_2^2 \leq (s_v(d_v - 1)).
$$

Note slice admission control decisions are clearly decoupled under SS, but paralleling Theorem 4 we have following result.

**Theorem 5.** Under Assumptions 1 and 3, the optimal admission control policy under SS are decoupled. The optimal choice for slice $v$ $\hat{\rho}^{v, SS}_*$ is the unique solution to:

$$
\min_{\tilde{\rho}^v \in \Gamma^v} \parallel \tilde{\rho}^v \parallel_2^2,
$$

and the associated carried load is given by $\rho^v_* = \frac{s_v d_v - 1}{\parallel \tilde{\rho}^{v, SS}_* \parallel_2^2}$.

By comparing Eq. (15) and (16), one can see that under SS, slices simply seek to balance their own relative loads on the network. By taking the ratio between $\rho^v_*$ and $\rho^v_{SS,*}$, one can show that under Assumptions 1, 2 and 3 the gain in carried load for slice $v$ is given by

$$
G^\text{load}_v \triangleq \frac{\rho^v_*}{\rho^v_{SS,*}} = \frac{\parallel \tilde{\rho}^{v, SS}_* \parallel_2^2}{\tilde{g}^v_* \parallel \tilde{\rho}^{v,*} \parallel_2^2} \times \frac{s_v (d_v - 1)}{s_v (d_v - 1) - 1}.
$$

The first factor captures a traffic shaping dependent gain for slice $v$. The second factor is a result of statistical multiplexing gains. A simple special case is highlighted in the following corollary.

**Corollary 2.** Under Assumptions 1, 2 and 3, if user mobility patterns are such that $\frac{1}{d_v} 1 \in \Gamma^v \forall v \in V$, the gain in the total carried load under the SCPF traffic shaping Nash equilibrium vs. optimal admission control for SS is given by:

$$
G^\text{load}_v = \frac{s_v d_v - s_v}{s_v d_v - 1} \geq 1, \forall v \in V.
$$

This result also follows directly from the KKT conditions associated with the admission control problem, and the observation that $\hat{\rho}^{v,*} = \frac{1}{d_v} 1, \forall v \in V$ at the Nash equilibrium. Then, substituting this solution into the BTD constraint one obtains the result for the gain in carried loads. The reader is referred to the extended version for a detailed proof [23].

Note that in order for a BTD constraint to be feasible under SS, one must require $s_v (d_v - 1) > 1$. It can be seen that the gain exhibited in Corollary 2 can be very high when $s_v \downarrow 1/d_v$. Furthermore if $s_v \uparrow 1$ we have that $G^\text{load} \downarrow 1$, i.e., no actual gain. This result implies that slices with small shares or tight
BTD constraints will benefit most from sharing, coinciding with our observations in Corollary 1.

V. PERFORMANCE EVALUATION

We simulated a wireless network shared by multiple slices supporting mobile customers following the IMT-Advanced evaluation guidelines [13]. The system consists of 19 base stations in a hexagonal cell layout with an intersite distance of 200 meters and 3 sector antennas, mimicking a dense ‘small cell’ deployment. Thus, in this system, \( B \) corresponds to 57 sectors. Users associate to the sector offering the strongest SINR, where the downlink SINR is modeled as in [22]:

\[
\text{SINR}_{ub} = \frac{P_b G_{ub}}{\sum_{k \in B, k \neq b} P_k G_{uk} + \sigma^2},
\]

where, following [13], the noise \( \sigma^2 = -10 \text{dB} \), the transmit power \( P_b = 41 \text{dB} \) and the channel gain between user \( u \) and BS sector \( b \), denoted by \( G_{ub} \), accounts for path loss, shadowing, fast fading and antenna gain. Letting \( d_{u,b} \) denote the current distance in meters from the user \( u \) to sector \( b \), the path loss is defined as \( 36.7 \log_{10}(d_{u,b}) + 22.7 + 26 \log_{10}(f_c) \text{dB} \), for a carrier frequency \( f_c = 2.5 \text{GHz} \). The antenna gain is set to 17 dBi, shadowing is updated every second and modeled by a log-normal distribution with standard deviation of 8dB, as in [22]; and fast fading follows a Rayleigh distribution depending on the mobile’s speed and the angle of incidence. The downlink rate \( c_u \) currently achievable to user \( u \) is based on discrete set modulation and coding schemes (MCS) and associated SINR thresholds given in [1]. This MCS value is selected based on the averaged \( \text{SINR}_{ub} \), where channel fast fading is averaged over a second.

We model slices’ with different spatial loads by modeling different customer mobility patterns. Roughly uniform spatial loads are obtained by simulating the Random Waypoint model [12], while non-uniform loads obtained by simulating the SLAW model [16]. These mobility models would not induce Markovian motion amongst base stations assumed in our analysis, yet the analytical results are robust to these assumptions.

A. Statistical Multiplexing and BTD Gains

We evaluated the BTD gains of SCPF vs SS for four simulation scenarios, each including 4 slices, each with equal shares but different spatial load patterns. For each scenario, we provide results for simulated BTD gains, and results from our theoretical analysis (Corollary 1) based on the empirically obtained spatial traffic loads. More detailed information regarding simulated scenarios and resulting empirical spatial traffic loads for high load regime are displayed in Table I and a snapshot of locations for the 4 slices’ users in a network with a load of 4 users per sector is displayed in Figure 1.

The results given in Figure 2 show the BTD gains for each scenario as the overall network load increases. In Scenario 3, the aggregate network traffic is “smoother” than the individual slice’s traffic, and the gains are indeed higher. This is also the case for Slice 1 and 2 in Scenario 4, since these slices loads are more “imbalanced” than the other two slices, they experience higher gains. In Scenario 2, where slices non-homogenous spatial loads are ‘aligned’, aggregation does not lead to smoothing and the gains are least.

As can be seen in Figure 2 the simulated and theoretical gains (dashed lines) of Corollary 1 are an excellent match. The theoretical model has been calibrated to the mean reciprocal capacities seen by slice customers (i.e., \( \gamma^b_v \)’s) and the measured induced loads resulting from the slice mobility patterns.

B. Traffic Shaping Equilibrium and Carried Load Gains

In order to study the equilibria reached by the traffic shaping game, we measured the underlying user mobility patterns in Section V-A, and modeled it via a random routing matrix. We further assumed uniform intensity of arrivals rates at all base stations and uniform exit probabilities of 0.1. The mean holding time at each base station was again calibrated with the simulations in Section V-A. We considered a traffic

| Scenario: Slices | Spatial loads | \( ||\bar{\rho}||_2 \) | \( ||\bar{g}||_2 \) | \( \theta(\bar{g}, \bar{\rho}) \) | \( G_{\rho \nu}^b \) |
|------------------|---------------|-------------------|---------------|-------------------|-------------|
| 1 Homogeneous uniform | 0.27 | 0.27 | 7.09 | 1.0 % |
| 2 Homogeneous non-uniform | 0.32 | 0.32 | 6.18 | 1.0 % |
| 3 Heterogeneous orthogonal | 0.36 | 0.26 | 41.78 | 83.3 % |
| 4 Mixed Slices 1&2 non-uniform | 0.36 | 0.23 | 25.52 | 70.4 % |
| 3&4 uniform | 0.19 | 0.23 | 48.00 | 23.7 % |

Table I: Measured normalized slice and network traffic norms and angles for highest load case of each scenario.
shaping game for a network shared by 3 slices, where Slice 1 has uniform spatial loads and Slice 2 and 3 have different non-uniform spatial loads. All slices have equal shares and their capacity normalized BTD requirements are set to $\tilde{d}_1 = 10$, $\tilde{d}_2 = 12$, $\tilde{d}_3 = 15$ respectively. The Nash equilibrium was solved via the algorithm included in [23]. The convergence is reached within 3 rounds of iterations under the parameters given in [23]. The results shown in Figure 3 exhibit dashed lines corresponding to the theoretical carried load gains in the saturated regime. As can be seen, these coincide with the Nash equilibria of the simulated traffic shaping games for high arrival rates. For lower arrival rates the gains can be much higher, e.g., almost 1.6x, for slices with non-uniform mobility patterns. This was to be expected since for lower loads we expect higher statistical multiplexing gains from sharing, and thus relatively higher carried loads to be admitted. For very low loads, as expected, there are no gains since all traffic can be admitted and BTD constraints are met.

Also shown in Figure 3(subfigure) is the degree to which the relative loads of slices $\tilde{\rho}^S$, and the weighted aggregate traffic on the network $\tilde{g}$ are balanced, as measured by $||\cdot||_2$, as the arrival rates on the network increase. As expected, based on Theorem 4, as arrivals increase relative loads of slices and the network become more balanced, showing the compromise the traffic shaping game is making, balancing slices relative loads and that of the overall network.

VI. Conclusions

This paper has thoroughly explored a relatively simple and natural approach for resource sharing amongst network slices – SCPF – which corresponds to socially optimal allocations in a Fisher market. Our analysis of performance in settings where slices support stochastic loads provides explicit formulas for (i) the performance gains one can expect over static slicing, (ii) how to dimension slice shares to meet performance objectives, and (iii) how to go about performance management through admission control. If dynamic resource sharing amongst network slices is to be adopted, the ability to realize disciplined engineering and performance prediction will be the key. Our analysis of SCPF seems to meet these requirements and at the same time reveals some intriguing insights regarding the load interactions in such sharing models, in particular the impact of relative load distributions on statistical multiplexing, and the role of traffic shaping in optimizing admission control. Finally, we note that our approach to admission control in an SCPF shared system is novel in that each slice exploits knowledge of its customers’ mobility patterns to optimize its carried load and assure service continuity.

REFERENCES