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 re-distributes 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

978-3-901882-90-6/17

(as a proxy for proportional fairness). Our focus here is on SCPF-based sharing amongst slices with stochastic loads and on "inelastic" or "rate-adaptive" customers, e.g., video, voice, and more generally customers on properly provisioned networks, whose activity on the network can be assumed to be independent of their resource allocations.

Finally there is much ongoing work on developing the network slicing concept, see e.g., [18], [24] and references therein, including development of approaches to network virtualization in RAN architectures, e.g, [8], [15], and SDN-based implementation, e.g., [4]. This paper focuses on devising good slice-based resource sharing criteria to be incorporated into such architectures.

Contributions of this paper. This paper makes several contributions centering on a simple and practical resource sharing mechanism: SCPF. First, we consider user performance (bit transmission delay) on slices supporting stochastic loads. In particular we develop expressions for (i) the mean performance seen by a typical user on a network slice; and (ii) the achievable performance gains versus static slicing. We show that when a slice's load is more 'imbalanced' than, and/or 'orthogonal' to, the aggregate network load, one will see higher performance gains. Our analysis provides an insightful picture of the "geometry" of statistical multiplexing for SCPF-based network slicing. Second, under SCPF, traditional network dimensioning translates to a coupled share dimensioning problem, which addresses whether there exist feasible share allocations given slices' expected loads and performance requirements. We provide a solution to robust share dimensioning for SCPF-based network slicing. Third, we consider decentralized per-slice performance management under SCPF sharing. In particular, we consider admission control aimed at maximizing a slice's carried load subject to a performance constraint. When slice unilaterally optimize their admission control policies, the coupling of their decisions can be viewed as a "traffic shaping" game, which is shown to have a Nash equilibrium. For a high load regime we explicitly characterize the equilibrium and the associated gains in carried load for SCPF vs static slicing. Finally, we present detailed simulations for a shared distributed infrastructure supporting slices with different mobility patterns which match our analysis well, and further support our conclusions on gains in both performance and carried loads of SCPF sharing.

II. System Model

A. Network Slices, Resources and Mobile Service Traffic

We consider a collection of base stations (sectors) $\mathcal{B}$ shared by a set of network slices $\mathcal{V}$, with cardinalities B and V respectively. For example, $\mathcal{V}$ might denote slices supporting different services or (virtual) mobile operators etc.

We envisage each slice v as supporting a mobile service in the region served by the base stations $\mathcal{B}$. Each slice supports a stochastic load of users (devices/customers) with an associated mobility/handoff policy. In particular, we assume that exogenous arrivals to slice v at base station b follow a Poisson process with intensity γ_b^v and $\boldsymbol{\gamma}^v = (\gamma_b^v : b \in \mathcal{B})^T$. Each slice v customer at base station b has an independent sojourn time with mean μ_b^v 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 $\boldsymbol{Q}^v = (q_{i,j}^v : i, j \in \mathcal{B})$ denote a slice-dependent routing matrix where $q_{i,j}^v$ is the probability a slice v customer moves from base station i to j and $1 - \sum_{j \in \mathcal{B}} q_{i,j}^v$ is the probability it exits the system. Throughout the paper, we assume $\boldsymbol{Q^v}$ is irreducible for all $v \in \mathcal{V}$. This model induces an overall traffic intensity for slice v across base stations satisfying flow conservation equations: for all $b \in \mathcal{B}$ we have

$$\kappa_b^v = \gamma_b^v + \sum_{a \in \mathcal{B}} \kappa_a^v q_{a,b}^v,$$

where κ_b^v 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 $\rho_b^v = \kappa_b^v \mu_b^v$, and $\boldsymbol{\rho}^v \triangleq (\rho_b^v : b \in \mathcal{B})^T$ captures its overall system load. Letting $\boldsymbol{\mu}^v = (\mu_b^v : b \in \mathcal{B})^T$, the flow conservation equations can be rewritten in matrix form as:

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

Note that $\boldsymbol{I} - (\boldsymbol{Q}^v)^T$ 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_b^v are mutually independent and $N_b^v \sim \text{Poisson}(\rho_b^v)$. Since the sum of independent Poisson random variables is again Poisson, the total number of customers on slice v is such that $N^v = \sum_{b \in \mathcal{B}} N_b^v \sim \text{Poisson}(\rho_v)$ where $\rho_v = \sum_{b \in \mathcal{B}} \rho_b^v$.

Our network model for the numbers of customers and mobility across base stations, assumes customer sojourns/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 $\mathcal{U}_b^v$,$\mathcal{U}_b$,$\mathcal{U}^v$ and $\mathcal{U}$ 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., $|\mathcal{U}_b^v| = n_b^v$ and $|\mathcal{U}^v| = n^v$, where in a stationary regime n^v and n_b^v are realizations of Poisson random variables N^v and N_b^v, respectively.

Each base station b is modeled as a finite resource shared by its associated users $\mathcal{U}_b$. A customer $u \in \mathcal{U}_b$ 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 $\sum_{u\in\mathcal{U}_b} f_u = 1$. 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 for a typical customer on slice v at base station b is an independent realization of a random variable with the same distribution as C_b^v. 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 slices is allocated a 'share' of the network resources $s_v, v \in \mathcal{V}$ such that $s_v > 0$ and $\sum_{v\in\mathcal{V}} s_v = 1$.

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 $u \in \mathcal{U}_b^v$ gets an equal share, i.e., $1/n_b^v$, of the slice v's resources at base station b. Thus the users transmission rate r_u^{SS} is given by*

$$r_u^{SS} = \frac{s_v}{n_b^v} c_u.$$

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) $w_u = \frac{s_v}{n_v}$ for $u \in \mathcal{U}^v, \forall v \in \mathcal{V}$. In turn, each base station allocates resources to customers in proportion to their weights. So a user $u \in \mathcal{U}_b^v$ gets a transmission rate r_u^{SCPF} given by*

$$r_u^{SCPF} = \frac{w_u}{\sum_{u'\in\mathcal{U}_b} w_{u'}} c_u = \frac{\frac{s_v}{n^v}}{\sum_{v'\in\mathcal{V}} \frac{n_b^{v'} s_{v'}}{n^{v'}}} c_u. \tag{2}$$

Thus under SCPF the *overall* fraction of resources slice v is allocated at a base station b is proportional to $\frac{n_b^v}{n^v} s_v$, 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 on 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_b^v, 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 $\mathbb{E}^v$ 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_b^v that of such customer of slice v at base station b. Similarly, let $R^{v,SS}$ and $R_b^{v,SS}$ denote these quantities under static slicing. Thus, under SCPF the average BTD for a typical slice v customer is given by $\mathbb{E}^v[\frac{1}{R^v}]$. 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 $\tilde{\boldsymbol{\rho}}^v = (\tilde{\rho}_b^v : b \in \mathcal{B})^T$ where $\tilde{\rho}_b^v \triangleq \frac{\rho_b^v}{\rho_v}$; the overall *share weighted normalized load distribution* is $\tilde{\boldsymbol{g}} = (\tilde{g}_b : b \in \mathcal{B})^T$ where $\tilde{g}_b \triangleq \sum_{v\in\mathcal{V}} s_v \tilde{\rho}_b^v$; and the mean reciprocal resource capacity for slice v is $\boldsymbol{\delta}^v = (\delta_b^v : b \in \mathcal{B})^T$ where $\delta_b^v \triangleq \mathbb{E}^v[\frac{1}{C_b^v}]$.

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

$$\mathbb{E}^v\left[\frac{1}{R^v}\right] = \sum_{b\in\mathcal{B}} \tilde{\rho}_b^v \delta_b^v \left(1 - \tilde{\rho}_b^v + \frac{(\rho_v + 1)}{s_v}\tilde{g}_b\right). \tag{3}$$

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

$$\mathbb{E}^v\left[\frac{1}{R^{v,SS}}\right] = \sum_{b\in\mathcal{B}} \tilde{\rho}_b^v \delta_b^v \left(\frac{\rho_b^v + 1}{s_v}\right). \tag{4}$$

Proof. Recall that Poisson arrivals see time averages, i.e., see the remaining users in the product-form stationary distribution, given in Section II-A. Thus the distribution as seen by a typical user on slice v at base station b is the same as the product-form distribution *plus an additional customer* on slice v at base station b. Using this fact and SCPF resource allocations as given by Eq. (2), the BTD of a typical slice v user at base station b can be expressed as follows:

$$\begin{aligned}\mathbb{E}^v\left[\frac{1}{R_b^v}\right] &= \mathbb{E}^v\left[\frac{1}{C_b^v}\right]\mathbb{E}\left[\frac{s_v\frac{N_b^v+1}{N^v+1}+\sum_{v'\neq v}\frac{s_{v'}N_b^{v'}}{N^{v'}}}{\frac{s_v}{(N^v+1)}}\right]\\ &= \delta_b^v\mathbb{E}\left[(N_b^v+1)+\frac{N^v+1}{s_v}\sum_{v'\neq v}\frac{s_{v'}N_b^{v'}}{N^{v'}}\right]\\ &= \delta_b^v\left(1-\tilde{\rho}_b^v+\frac{(\rho_v+1)}{s_v}\tilde{g}_b\right).\end{aligned}$$

Where the last equality follows by noticing that (i) N^v is independent of $N_b^{v'}$ and $N^{v'}$ and (ii) $E[\frac{N_b^{v'}}{N^{v'}}]=\frac{\rho_b^{v'}}{\rho^{v'}}$. 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(λ), then by symmetry we have

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

Under static slicing we have that

$$\mathbb{E}^v\left[\frac{1}{R_b^{v,SS}}\right]=\mathbb{E}^v\left[\frac{1}{C_b^v}\right]\mathbb{E}\left[\frac{(N_b^v+1)}{s_v}\right]=\delta_b^v\frac{\rho_b^v+1}{s_v}.$$

The theorem follows by taking an weighted average across base stations – weighted by the fraction of customers at each base station, i.e., $\tilde{\rho}_b^v$. $\square$

B. Analysis of Gain

Using the results in Theorem 1 one can evaluate the gains in the mean BTD for a typical slice v user under SCPF versus SS, i.e.,

$$G_v=\frac{\mathbb{E}^v\left[\frac{1}{R^{v,SS}}\right]}{\mathbb{E}^v\left[\frac{1}{R^v}\right]}.$$

In general, one would expect $G_v \geq 1$ since under SCPF typical users should see higher allocated rates and thus lower BTDs. One can verify that is the case when slices have *uniform* loads across base stations but the general case is more subtle. For simplicity from here on in this paper we focus on the case where the following additional assumption is in effect:

Assumption 1. *Base stations are said to be homogeneous for slice v if for all $b\in\mathcal{B}$: $\mathbb{E}^v\left[\frac{1}{C_b^v}\right]=\delta_v$.*

Note that Assumption 1 only requires the *average* reciprocal capacity a given slices' customer sees across base stations is homogenous.

Corollary 1. *Under Assumption 1, the BTD gain of SCPF over SS for slice v is given by*

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

For fixed relative loads $\tilde{\boldsymbol{\rho}}^v$ and $\tilde{\boldsymbol{g}}$, the gain for low overall load ($\rho_v\to 0$) is positive and given by:

$$G_v^L=\frac{1}{(\tilde{\boldsymbol{g}}-s_v\tilde{\boldsymbol{\rho}}^v)^T\tilde{\boldsymbol{\rho}}^v+s_v}\geq 1,$$

where $\tilde{\boldsymbol{g}}-s_v\tilde{\boldsymbol{\rho}}^v$ is the shared weighted relative load of other slices on network. Furthermore, G_v is a nonincreasing function of ρ_v, and for high loads ($\rho_v\to\infty$) is given by:

$$G_v^H=\frac{\|\tilde{\boldsymbol{\rho}}^v\|_2}{\|\tilde{\boldsymbol{g}}\|_2}\times\frac{1}{\cos(\theta(\tilde{\boldsymbol{g}},\tilde{\boldsymbol{\rho}}^v))},$$

where $\theta(\tilde{\boldsymbol{g}},\tilde{\boldsymbol{\rho}}^v)$ denotes the angle between the slice's relative loads and the overall share weighted relative loads on the network.

A detailed proof of Corollary 1 can be found in [23]. The result indicates that for slice v with fixed relative loads $\tilde{\rho}^v$, the gains decrease in the overall load ρ_v, thus if $G_v^H>1$ SCPF always provides a gain. A sufficient condition for gains under high loads is that $\|\tilde{\boldsymbol{g}}\|_2\leq\|\tilde{\boldsymbol{\rho}}^v\|_2$. Since $\|\tilde{\boldsymbol{g}}\|_1=\|\tilde{\boldsymbol{\rho}}^v\|_1=1$, this follows when the overall share weighted relative load on the network is more balanced than that of slice v. One would typically expect aggregated traffic to be more balanced than that of individual slices. This condition is fairly weak, i.e., it does not depend on where the loads are placed, but on how balanced they are. The corollary also suggests gains are higher when $\cos(\theta(\tilde{\boldsymbol{g}},\tilde{\boldsymbol{\rho}}^v))$ is small. In other words, a slice with imbalanced normalized loads whose relative load distribution is 'orthogonal' to the shared weighted aggregate traffic, i.e., $\theta(\tilde{\mathbf{g}},\tilde{\rho}^v)\approx 0$, will tend to see higher gains. The simulations in Section V further explore these observations.

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 $\tilde{\boldsymbol{g}}=\tilde{\boldsymbol{\rho}}^v$ and let $\tilde{d}_v\triangleq d_v/\delta_v$ denote slice v's normalized BTD constraint. Note that δ_v is the minimum BTD achievable when a user gets **all** the base station resources, so a target requirement satisfies $d_v>\delta_v$ and so $\tilde{d}_v>1$. For slice v to meet a mean BTD constraint $\tilde{d}_v$, it follows from Eq. (3) that :

$$\rho_v\leq l(\tilde{d}_v,\tilde{\boldsymbol{\rho}}^v)\triangleq\frac{\tilde{d}_v-1}{\|\tilde{\boldsymbol{\rho}}^v\|_2^2}.$$

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

$$s_v\geq\frac{1+\rho_v}{l(\tilde{d}_v,\tilde{\boldsymbol{\rho}}^v)-\rho_v}\sum_{u\neq v}s_u\frac{\|\tilde{\boldsymbol{\rho}}^u\|_2}{\|\tilde{\boldsymbol{\rho}}^v\|_2}\cos(\theta(\tilde{\boldsymbol{\rho}}^u,\tilde{\boldsymbol{\rho}}^v)).\quad(5)$$

This can be written as:

$$\sum_{v\in\mathcal{V}} s_v \boldsymbol{h}^v \succeq 0, \tag{6}$$

where we refer to $\boldsymbol{h}^v = (h_u^v : u \in \mathcal{V})^T$ as v's *share coupling vector*, given by

$$h_u^v = \begin{cases} 1 & v = u \\ -\frac{1+\rho_u}{l(\tilde{d}_u,\tilde{\boldsymbol{\rho}}^u)-\rho_u}\frac{\|\tilde{\boldsymbol{\rho}}^v\|_2}{\|\tilde{\boldsymbol{\rho}}^u\|_2}\cos(\theta(\tilde{\boldsymbol{\rho}}^u,\tilde{\boldsymbol{\rho}}^v)) & v \neq u. \end{cases}$$

We can interpret $h_v^v = 1$ as the benefit to slice v of allocating unit share to it. When $v \neq u$, h_u^v depends on two factors. The first $\frac{1+\rho^u}{l(\tilde{d}_u,\tilde{\boldsymbol{\rho}}^u)-\rho^u}$ captures the sensitivity of slice u to the 'share weighted congestion' from other slices. If ρ_u is close to its limit $l(\tilde{d}_u, \tilde{\boldsymbol{\rho}}^u)$, its sensitivity is naturally very high. The second term, $\frac{\|\tilde{\boldsymbol{\rho}}^v\|_2}{\|\tilde{\boldsymbol{\rho}}^u\|_2}\cos(\theta(\tilde{\boldsymbol{\rho}}^u,\tilde{\boldsymbol{\rho}}^v))$ 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 $((\rho_v, \tilde{\boldsymbol{\rho}}^v, d_v) : v \in \mathcal{V})$ are admissible under SCPF sharing if and only if there exists an $\boldsymbol{s} = (s_v : v \in \mathcal{V})^T$ such that $\|\boldsymbol{s}\|_1 = 1$, $\boldsymbol{s} \succeq \boldsymbol{0}$ and*

$$\sum_{v\in\mathcal{V}} s_v \boldsymbol{h}^v \succeq 0.$$

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

$$\max_{\boldsymbol{s}\succeq\boldsymbol{0}}\{ \min_i \sum_{v\in\mathcal{V}} s_v h_i^v : \|\boldsymbol{s}\|_1 = 1 \}. \tag{7}$$

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 $\boldsymbol{a}^v = (a_b^v : b \in \mathcal{B})^T \in [0,1]^B$ where a_b^v 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 $\gamma_b^v a_b^v$. Based on the flow conservation equation Eq. (1) one can obtain the carried load $\boldsymbol{\rho}^v$ induced by admission control policy $\boldsymbol{a}^v$ via

$$\boldsymbol{\rho}^v = (\boldsymbol{M}^v)^{-1}\boldsymbol{a}^v = \text{diag}(\boldsymbol{\mu}_v)(\boldsymbol{I}-(\boldsymbol{Q}^v)^T)^{-1}\text{diag}(\gamma_v)\boldsymbol{a}_v$$

where $\boldsymbol{M}^v \triangleq \text{diag}(\boldsymbol{\gamma}_v)^{-1}(\boldsymbol{I}-(\boldsymbol{Q}^v)^T)\text{diag}(\boldsymbol{\mu}_v)^{-1}$ is invertible because $\boldsymbol{I} - (\boldsymbol{Q}^v)^T$ is irreducibly diagonally dominant.[1] By contrast with Section II-A, note that $\boldsymbol{\rho}^v$ 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 $\tilde{\boldsymbol{g}}$ 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 SCPF based sharing.

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

$$\max_{\tilde{\boldsymbol{\rho}}^v,\rho_v} \quad \rho_v \tag{8}$$

$$\text{s.t.} \quad \boldsymbol{a}^v = \rho_v \boldsymbol{M}^v \tilde{\boldsymbol{\rho}}^v, \quad \boldsymbol{a}^v \in [0,1]^B, \quad \boldsymbol{1}^T\tilde{\boldsymbol{\rho}}^v = 1 \tag{9}$$

$$\frac{(\rho_v+1)}{s_v}\tilde{\boldsymbol{g}}^T\tilde{\boldsymbol{\rho}}^v - \|\tilde{\boldsymbol{\rho}}^v\|_2^2 \leq \tilde{d}_v - 1. \tag{10}$$

Note that Eq. (9) establishes a one-to-one mapping between $(\tilde{\boldsymbol{\rho}}^v, \rho_v)$ and $\boldsymbol{a}^v$. We will use $\tilde{\boldsymbol{\rho}}^v$ and ρ_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 on the network $\tilde{\boldsymbol{g}}$, the slice's load and its customer mobility patterns (i.e., $\boldsymbol{M}^v$). Unfortunately this problem is not convex due to the BTD constraint Eq. (10); however, for high overall per slice loads it is easily shown to be convex, hence for simplicity we will make the following additional assumption.

Assumption 2. *The network is said to see high overall slice loads, if for all $v \in \mathcal{V}$ 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:

$$\frac{(\rho_v+1)}{s_v}\tilde{\boldsymbol{g}}^T\tilde{\boldsymbol{\rho}}^v - \|\tilde{\boldsymbol{\rho}}^v\|_2^2 \approx \frac{\rho_v}{s_v}\tilde{\boldsymbol{g}}^T\tilde{\boldsymbol{\rho}}^v = (s_v x_v)^{-1}\tilde{\boldsymbol{g}}^T\tilde{\boldsymbol{\rho}}^v \tag{11}$$

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

$$f_v(\tilde{\boldsymbol{\rho}}^v;\tilde{\boldsymbol{\rho}}^{-v}) \triangleq \tilde{\boldsymbol{g}}^T\tilde{\boldsymbol{\rho}}^v \leq s_v(\tilde{d}_v - 1)x_v. \tag{12}$$

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

[1] If $\boldsymbol{\gamma}^v$ 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 $\boldsymbol{y}^{-v}$, slice v determines its admission control policy $\boldsymbol{y}^v = (\tilde{\boldsymbol{\rho}}^v, x^v)$ by solving

$$\min_{\boldsymbol{y}^v} \{ \; x_v \mid \boldsymbol{y}^v \in Y^v(\boldsymbol{y}^{-v}) \; \} \tag{13}$$

where $Y^v(\boldsymbol{y}^{-v})$ denotes slice v's feasible policies and is given by

$$\begin{aligned} Y^v(\boldsymbol{y}^{-v}) \;\; & \triangleq \{ \; \boldsymbol{y}^v \mid \mathbf{1}^T \tilde{\boldsymbol{\rho}}^v = 1, \; \mathbf{0} \preceq \boldsymbol{M}^v \tilde{\boldsymbol{\rho}}_v \preceq x_v \mathbf{1}, \\ & f_v(\tilde{\boldsymbol{\rho}}^v; \tilde{\boldsymbol{\rho}}^{-v}) \leq s_v(\tilde{d}_v - 1) x_v \; \}. \end{aligned} \tag{14}$$

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 $\mathbf{y}^* = (\mathbf{y}^{v,*}, v \in \mathcal{V})$ 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*[2] *in response to other slices' load is such that for all $v \in \mathcal{V}$, $\boldsymbol{a}^v \prec \mathbf{1}$.*

Assumption 3 depends on many factors including the BTD constraint, the mobility pattern and network slices' shares, but it is generally true when the exogenous traffic of all slices at all base stations γ_b^v is high. When 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 $\tilde{\boldsymbol{\rho}}^* \triangleq (\tilde{\boldsymbol{\rho}}^{v,*} : v \in \mathcal{V})$ are the unique solution to:*

$$\min_{(\tilde{\boldsymbol{\rho}}^v \in \Gamma^v : v \in \mathcal{V})} \| \sum_v s_v \tilde{\boldsymbol{\rho}}^v \|_2^2 + \sum_v s_v^2 \| \tilde{\boldsymbol{\rho}}^v \|_2^2, \tag{15}$$

where $\Gamma^v \triangleq \{ \; \tilde{\boldsymbol{\rho}}^v \mid \mathbf{1}^T \tilde{\boldsymbol{\rho}}^v = 1, \boldsymbol{M}^v \tilde{\boldsymbol{\rho}}^v \succeq \mathbf{0} \; \}$, and the associated carried load for slice v is $\rho_v^* = \frac{s_v(\tilde{d}_v - 1)}{\tilde{\boldsymbol{g}}^{*,T} \tilde{\boldsymbol{\rho}}^{v,*}}$, where $\tilde{\mathbf{g}}^*$ 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].

[2] Admission control under SS is defined in the sequel.

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 (ACSS$_v$): Under SS slice v can determine its optimal admission control $\mathbf{y}^v$ by solving:

$$\begin{aligned} \max_{\tilde{\boldsymbol{\rho}}^v, \rho_v} \quad & \rho_v \\ \text{s.t.} \quad & \boldsymbol{a}^v = \rho_v \boldsymbol{M}^v \tilde{\boldsymbol{\rho}}^v, \quad \boldsymbol{a}^v \in [0,1]^B \\ & \mathbf{1}^T \tilde{\boldsymbol{\rho}}^v = 1 \text{ and } \rho_v \| \tilde{\boldsymbol{\rho}}^v \|_2^2 \leq (s_v \tilde{d}_v - 1). \end{aligned}$$

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 $\tilde{\boldsymbol{\rho}}^{v,SS,*}$ is the unique solution to:*

$$\min_{\tilde{\boldsymbol{\rho}}^v \in \Gamma^v} \| \tilde{\boldsymbol{\rho}}^v \|_2^2, \tag{16}$$

and the associated carried load is given by $\rho_v^{SS,*} = \frac{s_v \tilde{d}_v - 1}{\| \tilde{\boldsymbol{\rho}}^{v,SS,*} \|_2^2}$.

By comparing Eq. (15) and Eq. (16), one can see that under SS, slices simply seek to balance their own relative loads on the network. By taking the ratio between ρ_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

$$G_v^{\text{load}} \triangleq \frac{\rho_v^*}{\rho_v^{SS,*}} = \frac{\| \tilde{\boldsymbol{\rho}}^{v,SS,*} \|_2^2}{\tilde{\boldsymbol{g}}^{*,T} \tilde{\boldsymbol{\rho}}^{v,*}} \times \frac{s_v(\tilde{d}_v - 1)}{s_v \tilde{d}_v - 1}. \tag{17}$$

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 $\frac{1}{B}\mathbf{1} \in \Gamma^v$ for all $v \in \mathcal{V}$, the gain in the total carried load under the SCPF traffic shaping Nash equilibrium vs. optimal admission control for SS is given by:*

$$G_v^{\text{load}} = \frac{s_v \tilde{d}_v - s_v}{s_v \tilde{d}_v - 1} \geq 1, \quad \forall v \in \mathcal{V}. \tag{18}$$

This result also follows directly from the KKT conditions associated with the admission control problem, and the observation that $\tilde{\boldsymbol{\rho}}^{v,*} = \frac{1}{B}\mathbf{1}, \forall v \in \mathcal{V}$ 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 $s_v \tilde{d}_v > 1$. It can be seen that the gain exhibited in Corollary 2 can be very high when $s_v \downarrow 1/\tilde{d}_v$. Furthermore if $s_v \uparrow 1$ we have that $G_v^{\text{load}} \downarrow 1$, 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, $\mathcal{B}$ corresponds to 57 sectors. Users associate to the sector offering the strongest SINR , where the downlink SINR is modeled as in [22]:

$$\mathrm{SINR}_{ub} = \frac{P_b G_{ub}}{\sum_{k\in\mathcal{B},k\neq b} P_k G_{uk} + \sigma^2},$$

where, following [13], the noise $\sigma^2 = -104$dB, the transmit power $P_b = 41$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 $36.7\log_{10}(d_{ub}) + 22.7 + 26\log_{10}(f_c)$dB, for a carrier frequency $f_c = 2.5$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 $\overline{\mathrm{SINR}}_{ub}$, 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

Scenario: Slices	Spatial loads	$\Vert\tilde{\boldsymbol{\rho}}^v\Vert_2$	$\Vert\tilde{\boldsymbol{g}}\Vert_2$	$\theta(\tilde{\boldsymbol{g}},\tilde{\boldsymbol{\rho}}^v)$	G_v^H
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.

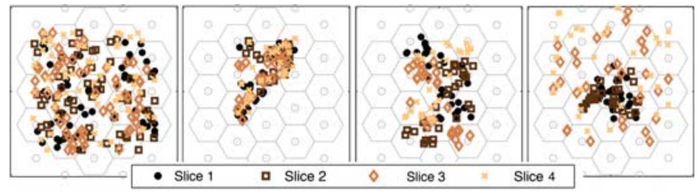

Fig. 1: Snapshot of users positions per slice and scenario exhibiting the different characteristics of traffic spatial loads. Left to right: Scenarios 1 to 4.

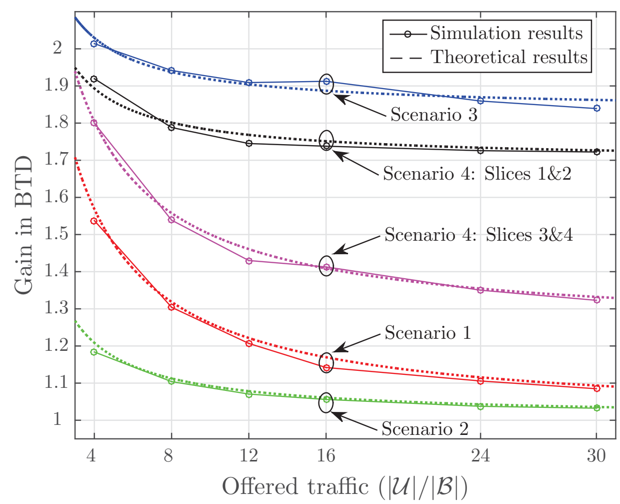

Fig. 2: BTD gain for our 4 different scenarios simulation.

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., γ_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

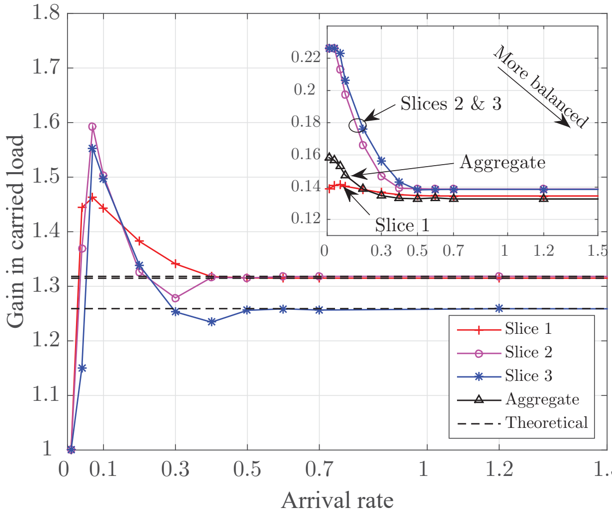

Fig. 3: Gain in carried load for various arrival rates. Subfigure: Balancing in relative load.

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 $\tilde{\boldsymbol{\rho}}^v$, and the weighted aggregate traffic on the network $\tilde{\boldsymbol{g}}$ 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

[1] Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures. 3GPP TS 36.213, v12.5.0, Rel. 12, Mar. 2015.

[2] F. Baccelli and P. Brémaud. *Palm probabilities and stationary queues*, volume 41. Springer Science & Business Media, 2012.

[3] A. Banchs. User fair queuing: fair allocation of bandwidth for users. In *IEEE INFOCOM*, volume 3, pages 1668–1677. IEEE, 2002.

[4] C. J. Bernardos et al. An architecture for software defined wireless networking. *IEEE Wireless Communications*, 21(3):52–61, 2014.

[5] T. Bonald and L. Massoulié. Impact of fairness on Internet performance. In *ACM SIGMETRICS*, volume 29, pages 82–91, 2001.

[6] T. Bonald and A. Proutiere. Insensitive bandwidth sharing in data networks. *Queueing systems*, 44(1):69–100, 2003.

[7] S. Brânzei, Y. Chen, X. Deng, A. Filos-Ratsikas, S. K. S. Frederiksen, and J. Zhang. The fisher market game: Equilibrium and welfare. 2014.

[8] X. Costa-Perez et al. Radio access network virtualization for future mobile carrier networks. *IEEE Comm. Magazine*, 51(7):27–35, 2013.

[9] G. de Veciana et al. Stability and performance analysis of networks supporting elastic services. *IEEE/ACM TON*, 9(1):2–14, 2001.

[10] C. Dutang. Existence theorems for generalized Nash equilibrium problems. *J. Nonlinear Analysis & Opt.*, 4(2):115–126, 2013.

[11] M. Feldman et al. The proportional-share allocation market for computational resources. *IEEE TPDS*, 20(8):1075–1088, 2009.

[12] E. Hyytia, P. Lassila, and J. Virtamo. Spatial node distribution of the random waypoint mobility model with applications. *IEEE TMC*, 5(6):680–694, June 2006.

[13] ITU-R. Report ITU-R M.2135-1, Guidelines for evaluation of radio interface technologies for IMT-Advanced. Technical Report, 2009.

[14] F. P. Kelly. *Reversibility and Stochastic Networks*. Cambridge University Press, 2011.

[15] R. Kokku et al. NVS: A Substrate for Virtualizing Wireless Resources in Cellular Networks. *IEEE/ACM TON*, 20(5):1333–1346, 2012.

[16] K. Lee et al. SLAW: Self-similar least-action human walk. *IEEE/ACM TON*, 20(2):515–529, 2012.

[17] N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani. *Algorithmic game theory*, volume 1. Cambridge University Press Cambridge, 2007.

[18] M. Richart, J. Baliosian, J. Serrat, and J. L. Gorricho. Resource slicing in virtual wireless networks: A survey. *IEEE Transactions on Network and Service Management*, 13(3):462–476, Sept 2016.

[19] V. Shah and G. de Veciana. Performance evaluation and asymptotics for content delivery networks. In *IEEE INFOCOM*, 2014.

[20] R. Srikant and L. Ying. *Communication networks: an optimization, control, and stochastic networks perspective*. Cambridge University Press, 2013.

[21] S. J. Yang and G. de Veciana. Enhancing both network and user performance for networks supporting best effort traffic. *IEEE/ACM TON*, 12(2):349–360, 2004.

[22] Q. Ye et al. User Association for Load Balancing in Heterogeneous Cellular Networks. *IEEE Trans. Wireless Comm.*, pages 2706–2716, 2013.

[23] J. Zheng, P. Caballero, G. de Veciana, S. Baek, and A. Banchs. Optimizing statistical multiplexing of network slices on shared infrastructure. https://www.dropbox.com/s/3vx2m33efjghkjj/extended-paper-long.pdf.

[24] X. Zhou, R. Li, T. Chen, and H. Zhang. Network slicing as a service: enabling enterprises' own software-defined cellular networks. *IEEE Communications Magazine*, 54(7):146–153, Jul. 2016.