Sharing Within Limits: Partial Resource Pooling in Loss Systems

Anvitha Nandigam, Suraj Jog, D. Manjunath, Jayakrishnan Nair, and Balakrishna J. Prabhu

***Abstract*— Fragmentation of expensive resources, e.g., the spectrum for wireless services, between providers can introduce inefficiencies in resource utilization and worsen overall system performance. In such cases, resource pooling between independent service providers can be used to improve performance. However, for providers to agree to pool their resources, the arrangement has to be mutually beneficial. The traditional notion of resource pooling, which implies complete sharing, need not have this property. For example, under full pooling, one of the providers may be worse off and hence has no incentive to participate. In this paper, we propose *partial* resource sharing models as a generalization of full pooling, which can be configured to be beneficial to all participants. We formally define and analyze two partial sharing models between two service providers, each of which is an Erlang-B loss system with the blocking probabilities as the performance measure. We show that there always exist partial sharing configurations that are beneficial to both providers, irrespective of the load and the number of circuits of each of the providers. A key result is that the Pareto frontier has at least one of the providers sharing all its resources with the other. Furthermore, full pooling may not lie inside this Pareto set. The choice of the sharing configurations within the Pareto set is formalized based on the bargaining theory. Finally, large system approximations of the blocking probabilities in the quality-efficiency-driven regime are presented.**

***Index Terms*— Resource pooling, partial pooling, loss systems, spectrum sharing, Pareto frontier, bargaining theory.**

I. Introduction

HIGH availability is an important requirement of many services like wireless communications, cloud computing, hospitals, and fire fighting services. The resources required to provide these are expensive—think spectrum and base stations for wireless communication, servers and associated infrastructure for cloud computing, medical equipment and doctors for hospitals, fire trucks and trained personnel for fire fighting services. Service denial, which is the inability of the resources to satisfactorily meet a fraction of the demand, is an important performance measure for these services. When the demand is stochastic, the amount of resources required to provide a prescribed grade of service may be such that the utilization is low, especially in smaller systems. This means that small providers require more resources for a given service level. This in turn can make these services expensive for small providers. However, large systems experience statistical multiplexing gains and hence achieve economies of scale. Thus resource sharing or resource pooling can be useful when there are several independent entities providing similar services using similar resources.

Typically, resource pooling is assumed to involve the combining of the resources of all the participating providers and treating the combined system as one unit. In this paper we propose *partial resource pooling* as a generalization of the full pooling models. Specifically, we consider two loss systems modeled as M/M/N/N queues that operate independently in that they manage their own calls but they cooperate by pooling their servers partially as follows. When an overflow call arrives at one of the systems, (i.e., the number of active calls of the provider is greater than the number of servers it has), the other provider *may* loan one of its free servers in which case the call will be admitted. The server is loaned for the duration of the call. The overflow call is lost if the other provider chooses not to loan the server. The partial sharing model determines when such an overflow call is admitted. At one extreme would be the no pooling case where all overflow calls are lost and at the other extreme is the full pooling case where all overflow calls are admitted if there is a free server.

As mentioned above, several resource pooling models are available in the literature with the key feature being independent service systems, managed by independent decision makers, cooperating *fully,* acting as a single entity, and sharing the costs and/or benefits suitably. In other words, it is an all-or-nothing game with the parties either pooling their resources completely or staying out of the coalition and operating on their own. These models typically use cooperative or coalitional game theoretic ideas to determine the answers to the following questions. (1) Which entities will form a cooperating unit? (2) How are the revenues and costs shared?

In [2], [3] independent wireless network operators share base station infrastructure and spectrum to efficiently serve their customers. Stable cost sharing arrangements between the network operators are explored in this setting. Note that the sharing model here involves complete pooling of the spectrum and the base stations, as opposed to the opportunistic sharing of resources with secondary users in cognitive radio systems. In the system studied in [4], the cooperating entities choose the quantity of resources to provide a specified service grade and stable cost sharing arrangements are determined.

Manuscript received August 20, 2018; revised March 12, 2019; accepted April 28, 2019; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor H. Jiang. This work was supported in part by DST, in part by CEFIPRA, and in part by the Bharti Center for Communication. A preliminary version of this work appeared in the proceedings of COMSNETS 2016 [1]. *(Corresponding author: Jayakrishnan Nair.)*

A. Nandigam was with the Department of Electrical Engineering, IIT Bombay, Mumbai 400076, India.

S. Jog is with the ECE Department, University of Illinois Urbana–Champaign, Champaign, IL 61801 USA.

D. Manjunath and J. Nair are with the Department of Electrical Engineering, IIT Bombay, Mumbai 400076, India (e-mail: jayakrishnan.nair@ee.iitb.ac.in).

B. J. Prabhu is with LAAS-CNRS, Université de Toulouse, CNRS, 31059 Toulouse, France.

Digital Object Identifier 10.1109/TNET.2019.2918164

Server pooling has also been studied in the context of reengineering of manufacturing lines by modeling them as Jackson networks. Here several service stations are combined into one service station that is capable of providing the services of all the components, e.g., [5] and references therein.

More abstract forms of resource pooling have also been considered in the queuing literature. In [6], cooperating single server queues are combined into one single server queue whose service rate is upper bounded by the sum of the capacities. The actual service rate is determined by a cost structure and the service grade. In [7], cooperation among queues to optimally invest in a common service capacity, or choose the optimal demand to serve as a common entity, is analyzed.

To motivate our break from the preceding literature, consider the following example of two M/M/N/N loss systems, a standard model used in the dimensioning of cellular telephone systems. Provider P_1, with 85 channels and a load of 88 Erlangs, has a blocking probability of 0.1 and Provider P_2, with 59 channels and 70 Erlangs load, has a blocking probability of 0.2. If the two providers are combined into one, the joint system would have a combined load of 158 Erlangs served by 144 channels with blocking probability of 0.11. Clearly cooperation is beneficial to P_2, but unacceptable to P_1. And if blocking probability were the only performance measure, it is a case of "and never the twain shall meet." The partial pooling mechanisms that we develop in this paper allow *both* the operators to improve their performance.[1] To the best of our knowledge, such *partial* resource pooling with the guarantee of *individual rationality* has not been explored previously in the literature.

The rest of the paper is organized as follows. In the next section, we introduce the system model and describe two partial sharing models: the *bounded overflow* sharing model and the *probabilistic* sharing model. The blocking probabilities under these models and their monotonicity properties are also derived. In Section III, we characterize the *Pareto frontier* of the sharing configurations. The key result is that the Pareto frontier is non empty and is at the boundary of all possible sharing configurations—one of the providers has to always yield its free servers to overflow calls of the other. In Section IV, we characterize the economics of partial sharing by treating the sharing that emerges as the solution of Nash bargaining, Kalai-Smorodinsky bargaining, egalitarian sharing (both parties experience the same benefit) and utilitarian sharing (maximize the system benefit). The utility sets over which these bargaining solutions will be computed for our model do not satisfy the usual properties of convexity or comprehensiveness, making it less straightforward to guarantee the uniqueness of the bargaining solution. Nevertheless, using monotonicity properties of the blocking probabilities shown in Section III, we are able to show uniqueness of the Kalai-Smorodinsky and egalitarian solutions. Via numerical experiments, we demonstrate the contrasts between the different bargaining solutions, and also the potential benefits of partial resource pooling for both providers. In Section V, we address the computational complexity of the blocking probabilities for large loss systems [9]. We consider large system limits under the well known quality-efficiency-driven (QED) regime. Our large system analysis provides computationally light, yet accurate approximations of the blocking probabilities for realistic system settings. Finally, we conclude with a discussion on alternate sharing models, connections to more familiar models from the circuit multiplexing literature, alternate applications, and future work in Section VI.

[1]Indeed, an important motivation for this work comes from regulations around spectrum sharing that are now being developed, e.g., [8].

II. Model and Preliminaries

In this section, we describe our system model, propose our mechanisms for partial resource pooling, and state some preliminary results.

We begin by describing the baseline model with no resource pooling. We consider two service providers, P_1 and P_2. Each provider is modeled as an M/M/N/N queue or an Erlang-B loss system. Specifically, P_i has N_i servers/circuits. Calls arrive for service at P_i according to a Poisson process of rate λ_i. When a call arrives, it begins service at a free server if one is available. If all servers are busy, then the call is blocked. The holding times (a.k.a. service times) of calls at P_i are i.i.d., with S_i denoting a generic call holding time. We assume that $\mathbb{E}[S_i] =: \frac{1}{\mu_i} < \infty$. Thus, the offered load seen by P_i is given by $a_i := \frac{\lambda_i}{\mu_i}$. With no resource pooling between the providers, it is well known that the steady state call blocking probability for P_i is given by the Erlang-B formula:

$$E(N_i, a_i) = \frac{a_i^{N_i}}{N_i!} \left[\sum_{j=0}^{N_i} \frac{a_i^j}{j!} \right]^{-1}.$$

It is also well known that the steady state call blocking probability is insensitive to the *distribution* of the call holding times, i.e., it depends only on the *average* call holding time. Moreover, the blocking probability depends on the workload only through the offered load a_i.

Next, we describe the proposed partial resource pooling models.

A. Probabilistic Sharing Model

The probabilistic sharing model is parameterized by the tuple $(x_1, x_2) \in [0,1]^2$. Informally, under this model, P_i accepts an overflow call from P_{-i} with probability x_i.[2]

Formally, the probabilistic sharing model is defined as follows. Let n_i denote the number of active calls of P_i. When a call of P_{-i} arrives,

- If $n_{-i} < N_{-i}$, and $n_1 + n_2 < N_1 + N_2$, the call is admitted
- If $n_{-i} \geq N_{-i}$ and $n_1 + n_2 < N_1 + N_2$, the call is admitted with probability x_i
- If $n_1 + n_2 = N_1 + N_2$, the call is blocked

The vector $x := (x_1, x_2)$ defines the (partial) sharing configuration. Note that x_i captures the extent to which P_i pools its resources with P_{-i}. In particular, the configuration $(0,0)$ corresponds to no pooling, and the configuration $(1,1)$ corresponds to complete pooling. Moreover, note that the probabilistic sharing model does not keep track of whether an ongoing call of P_i is occupying a server of P_i or P_{-i}. This simplification, which makes the model analytically tractable, is identical to the *maximum packing* or *call repacking* model of [10], [11] and has been used extensively in the literature. One interpretation of this assumption is that once a P_i server

[2]When referring to the provider labeled i, we use $-i$ to refer to the other provider.

becomes free, if there are any ongoing P_i calls on P_{-i} servers, one of those is instantaneously shifted to the free P_i server.

Next, we characterize the steady state blocking probabilities under this partial sharing model. To do so, we define the following subsets of $\mathbb{Z}_+^2$. From hereon, the superscript p indicates probabilistic sharing.

$$\begin{aligned} M^{(p)} &:= \{(n_1,n_2):\ n_1+n_2 \le N_1+N_2\} \\ R^{(p)} &:= \{(n_1,n_2):\ n_1+n_2 = N_1+N_2\} \\ D_i^{(p)} &:= \{(n_1,n_2): n_i \ge N_i, n_1+n_2 < N_1+N_2\} \\ &\quad (i \in \{1,2\}). \end{aligned}$$

Here $M^{(p)}$ refers to the set of feasible states, $R^{(p)}$ corresponds to the feasible states when all the servers are busy, and $D_i^{(p)}$ are the states in which calls of P_i are accepted with probability x_{-i}.

Lemma 1: Under the probabilistic sharing model, the steady state blocking probability for Provider i is given by

$$B_i^{(p)}(x_1,x_2) = \frac{1}{G}\Bigg[\sum_{(n_1,n_2)\in R^{(p)}} f_1(n_1)f_2(n_2) + \sum_{(n_1,n_2)\in D_i^{(p)}} f_1(n_1)f_2(n_2)(1-x_{-i})\Bigg],$$

where

$$f_i(n) = \begin{cases} a_i^n/n! & \text{if } n < N_i \\ a_i^n x_{-i}^{n-N_i}/n! & \text{if } N_i \le n \le N_1+N_2, \end{cases}$$
$$G = \sum_{(n_1,n_2)\in M^{(p)}} f_1(n_1)f_2(n_2).$$

A key takeaway from Lemma 1 is that under the probabilistic sharing model, the steady state blocking probabilities remain insensitive to the distributions of the call holding times. Moreover, the dependence of the incoming workload on each provider's blocking probability is only through the vector of offered loads (a_1,a_2). Finally, note that

$$\begin{aligned} B_i^{(p)}(0,0) &= E(N_i,a_i), \\ B_1^{(p)}(N_1,N_2) &= B_2^{(p)}(N_1,N_2) = E(N_1+N_2,a_1+a_2). \end{aligned}$$

Proof of Lemma 1: Assuming that the call holding times are exponentially distributed, the state (n_1,n_2) of the system evolves as a continuous time Markov chain (CTMC) over M. It is easy to check that this CTMC is time-reversible and its invariant distribution π has a product form:

$$\pi(n_1,n_2) = \frac{f_1(n_1)f_2(n_2)}{G.}$$

The steady state blocking probability is then obtained by invoking the PASTA property. The insensitivity of the blocking probabilities to the call holding time distributions is a direct consequence of the reversibility of the above CTMC [12].

B. Bounded Overflow Pooling Model

The bounded overflow (BO) model is parameterized by the tuple (k_1,k_2), where $k_i \in [0,N_i]$. Informally, under the BO model, P_i accepts up to k_i overflow calls from the other provider P_{-i}. Thus, k_i is indicative of the extent to which P_i shares its resources with P_{-i}. We use randomization to let k_i take real values in $[0,N_i]$; specifically, P_i admits up to $\lfloor k_i \rfloor$ overflow calls from P_{-i}, and admits a $\lceil k_i \rceil$-th overflow call with probability $\{k_i\}$, where $\{k_i\} := k_i - \lfloor k_i \rfloor$ denotes the fractional part of k_i.

Formally, the BO model is defined as follows. Recall that n_i denotes the number of active calls of P_i. When a call of P_{-i} arrives,

- If $n_{-i} < N_{-i} + \lfloor k_i \rfloor$ and $n_1+n_2 < N_1+N_2$, the call is admitted
- If $n_{-i} = N_{-i} + \lfloor k_i \rfloor$ and $n_1+n_2 < N_1+N_2$, the call is admitted with probability $\{k_i\}$
- Else, the call is blocked

We refer to the tuple (k_1,k_2) as the (partial) sharing configuration between P_1 and P_2. Under the BO model, P_i can have at most $N_i + \lceil k_{-i} \rceil$ concurrent calls. Note that $(0,0)$ corresponds to no resource pooling and (N_1,N_2) corresponds to full pooling between the providers. Finally, we note that the BO model also assumes call repacking [10], [11].

Next, we characterize the blocking probability of each provider under the BO model; Henceforth, the superscript bo indicates the bounded overflow sharing model. To express the blocking probabilities, we define the following subsets of $\mathbb{Z}_+^2$.

$$M^{(bo)} := \left\{(n_1,n_2): \begin{smallmatrix} n_1 \le N_1+\lceil k_2 \rceil \\ n_2 \le N_2+\lceil k_1 \rceil \\ n_1+n_2 \le N_1+N_2 \end{smallmatrix}\right\}$$
$$R^{(bo)} := \left\{(n_1,n_2): \begin{smallmatrix} n_1 \le N_1+\lceil k_2 \rceil \\ n_2 \le N_2+\lceil k_1 \rceil \\ n_1+n_2 = N_1+N_2 \end{smallmatrix}\right\}$$

For $i \in \{1,2\}$,

$$\begin{aligned} C_i^{(bo)} &:= \{(n_1,n_2): n_i = N_i + \lceil k_{-i} \rceil,\ n_{-i} < N_{-i} - \lceil k_{-i} \rceil\}, \\ D_i^{(bo)} &:= \{(n_1,n_2): n_i = N_i + \lfloor k_{-i} \rfloor,\ n_{-i} < N_{-i} - \lfloor k_{-i} \rfloor\}. \end{aligned}$$

Here $M^{(bo)}$ refers to the set of feasible states, $R^{(bo)}$ corresponds to the feasible states when all the servers are busy, $C_i^{(bo)}$ is the set of feasible states for which arriving calls of P_i are blocked due to the constraint on the number of overflow calls, and $D_i^{(bo)}$ are the states for which calls of P_i are accepted with probability $\{k_{-i}\}$.

The following lemma characterizes the blocking probabilities of both providers under the BO partial sharing model.

Lemma 2: Under the bounded overflow sharing model, the steady state blocking probability for provider P_i is given by

$$\begin{aligned} &B_i^{(bo)}(k_1,k_2) \\ &= \frac{1}{G}\Bigg[\sum_{(n_1,n_2)\in R^{(bo)}\cup C_i^{(bo)}} f_1(n_1)f_2(n_2) \\ &\quad + \mathbf{1}_{[\{k_{-i}\}\neq 0]}(1-\{k_{-i}\}) \sum_{(n_1,n_2)\in D_i^{(bo)}} f_1(n_1)f_2(n_2)\Bigg], \end{aligned}$$

where

$$f_i(n) = \begin{cases} a_i^n/n! & \text{if } n \le N_i + \lfloor k_{-i} \rfloor \\ \{k_{-i}\}\, a_i^n/n! & \text{if } n = N_i + \lfloor k_{-i} \rfloor + 1, \end{cases}$$
$$G = \sum_{(n_1,n_2)\in M^{(bo)}} f_1(n_1)f_2(n_2).$$

As before, note that the steady state blocking probabilities are insensitive to the distributions of the call holding times, and

depend only on the vector of offered loads (a_1, a_2). Moreover,

$$B_i^{(bo)}(0,0) = E(N_i, a_i),$$
$$B_1^{(bo)}(N_1, N_2) = B_2^{(bo)}(N_1, N_2) = E(N_1 + N_2, a_1 + a_2).$$

The proof of Lemma 2, being similar to that of Lemma 1, is omitted.

It is easy to see that the blocking probabilities for both sharing models can be calculated in $O((N_1 + N_2)^2)$ steps where each step corresponds to calculating the product form expression $f_1(n)f_2(n)$ for a state n.

C. Monotonicity Properties of the Blocking Probabilities

We conclude this section by collecting some monotonicity properties of the blocking probabilities under the above partial sharing models. These properties play a key role in our analysis of the game theoretic aspects of partial sharing in Sections III and IV.

When stating results that apply to both sharing models, we refer to the steady state blocking probability of Provider i as $B_i(x_1, x_2)$, with the understanding that this represents

- $B_i^{(p)}(x_1, x_2)$ under the probabilistic sharing model,
- $B_i^{(bo)}(x_1\ N_1, x_2\ N_2)$ under the bounded overflow sharing model (i.e., $x_i = \frac{k_i}{N_i}$).

Note that the overall steady state blocking probability of the system is given by

$$B_{ov}(x_1, x_2) = \frac{\lambda_1}{\lambda_1 + \lambda_2} B_1(x_1, x_2) + \frac{\lambda_2}{\lambda_1 + \lambda_2} B_2(x_1, x_2).$$

Our monotonicity results are summarized in the following theorem.

Theorem 1: Under the probabilistic as well as the bounded overflow partial sharing models, the steady state blocking probabilities satisfy the following properties, for $i \in (1, 2)$.

1) *$B_i(x_1, x_2)$ is a strictly increasing function of x_i.*
2) *$B_{-i}(x_1, x_2)$ is a strictly decreasing function of x_i.*
3) *If $\mu_1 = \mu_2$, then $B_{ov}(x_1, x_2)$ is a strictly decreasing function of x_i.*

Theorem 1 highlights the impact of an increase in x_i on the blocking probabilities of P_i and P_{-i}, as well as the overall blocking probability. In particular, an increase in x_i (i.e., an increase in the extent to which P_i shares its servers with P_{-i}) decreases the fraction of blocked calls at P_{-i}, at the expense of increasing the fraction of blocked calls at P_i. Note that Statements 1 and 2 imply that $(0, 0)$ is the unique Nash equilibrium between the providers, assuming that the utility of each provider is a strictly decreasing function of its blocking probability. This means that a non-cooperative interaction sans signaling would not yield a mutually beneficial partial sharing configuration between the providers. In contrast, we show in Section IV that a *bargaining-based* interaction would indeed result in mutually beneficial partial sharing configurations.

Finally, Statement 3 of Theorem 1 highlights that so long as the mean call holding times are matched across both providers, an increase in x_i results in an overall reduction in the call drop probability of the system. This is because increasing x_i provides additional opportunities for calls to get admitted when there are free circuits. In particular, Statement 3 above implies that for $(x_1, x_2) \notin \{(0, 0), (1, 1)\}$,

$$B_{ov}(1, 1) < B_{ov}(x_1, x_2) < B_{ov}(0, 0),$$

implying that complete pooling minimizes the overall blocking probability of the system (when $\mu_1 = \mu_2$).

Note that even through the statement of Theorem 1 applies compactly to both sharing models, a separate proof is required for each model. We provide the proof of Theorem 1 for the bounded overflow sharing model in Appendix A. We omit the proof for the probabilistic sharing model due to space constraints; it can be found in Appendix D of the extended version of this paper [13].

out that while the statement of Theorem 1 seems intuitive, the proof is fairly non-trivial. In particular, our proof of Statement 3 for the bounded overflow model involves a subtle sample path argument (see Appendix A).

III. Efficient Partial Sharing Configurations

We have seen that complete resource pooling between providers is not necessarily stable, in the sense that it is not guaranteed to be beneficial to both providers. Having defined mechanisms for partial resource sharing in Section II, the natural questions that arise are:

1) Do there exist stable partial sharing configurations?
2) If so, can one characterize the Pareto frontier of the space of partial sharing configurations?

The goal of this section is to address the above questions.

First, we prove that under both the sharing mechanisms defined in Section II, there exist stable partial sharing configurations, i.e., there exist partial sharing configurations that result in a strictly lower blocking probability for each provider, compared to the case of no pooling. Next, we focus on characterizing the set of Pareto-efficient partial sharing configurations. Intuitively, this is the set of 'efficient' sharing configurations, over which it is not possible to lower the blocking probability for any provider without increasing the blocking probability of the other. Our main result is that any Pareto sharing configuration has at least one provider pooling all of its servers (i.e., $x_i = 1$ for some i).[3] Intuitively, efficient partial sharing configurations involve the more congested provider pooling all of its servers, enabling both providers to benefit from the resulting statistical economies of scale. Finally, we provide an exact characterization of the set of Pareto efficient sharing configurations (a.k.a. the Pareto frontier) under the probabilistic and bounded overflow partial sharing models.

We begin by defining 'stable' partial sharing configurations.

Definition 1: A sharing configuration (x_1, x_2) is QoS-stable *if $B_i(x_1, x_2) < E(N_i, a_i)$ for $i = 1, 2$.*

The following lemma guarantees the existence of QoS-stable sharing configurations.

Lemma 3: Under the probabilistic as well as the bounded overflow partial sharing models, the set of QoS-stable partial sharing configurations is non-empty.

Lemma 3 essentially validates our partial sharing mechanisms. Specifically, it asserts that even when the providers are highly asymmetric with respect to capacity and/or offered load, and even when complete resource pooling is not beneficial to one of the providers, there exists a partial sharing configuration that is beneficial to both providers. We omit the proof of Lemma 3 since it is a direct consequence of Lemma 4 below.

[3] P_i pooling all its servers means that it always yields a free server to an overflow call from P_{-i}.

Now that we are certain that mutually beneficial partial sharing configurations exist, we turn to the characterization of the set of efficient configurations. We begin by defining Pareto-efficient sharing configurations.

Definition 2: A sharing configuration (x_1, x_2) *is* Pareto-efficient *if*

1) (x_1, x_2) *is QoS-stable,*
2) *there does not exist a sharing configuration* (x_1', x_2') *such that* $B_i(x_1', x_2') \leq B_i(x_1, x_2)$ *for all* $i \in \{1, 2\}$ *and* $B_i(x_1', x_2') < B_i(x_1, x_2)$ *for some* $i \in \{1, 2\}$.

Condition 2 above is the standard definition of Pareto-efficiency—a configuration is Pareto-efficient if it is not possible to enhance the utility of one party (the utility of a provider being a strictly decreasing function of its blocking probability) without diminishing the utility of the other. Since our interest is in capturing the set of configurations that the providers could potentially agree upon, it is also natural to impose the requirement that each provider benefits from the partial sharing agreement; this is captured by Condition 1 in the definition.

Our main result is that at any Pareto-efficient sharing configuration, at least one provider pools all of its servers.

Theorem 2: Under the probabilistic as well as the bounded overflow partial sharing models, the set of Pareto-efficient sharing configurations is non-empty. Moreover, any Pareto-stable sharing configuration (x_1, x_2) *satisfies the property that* $x_i = 1$ *for some* $i \in \{1, 2\}$.

Intuitively, if the providers are symmetric with respect to offered load and number of servers, full pooling ($x_i = 1$ for all i) is Pareto-efficient, thanks to the statistical economies of scale in the pooled system. Theorem 2 shows that under general (possibly asymmetric) settings, where full pooling may not be QoS-stable, efficient configurations still involve at least one provider pooling all its servers. Indeed, statistical economies of scale lie at the heart of this result as well, as is highlighted by Lemma 4 stated below, which forms the basis of the proof of Theorem 2.

In stating Lemma 4 and proving Theorem 2, we use the following notation: $\mathcal{X} := [0, 1]^2, \quad \mathcal{X}^o := [0, 1)^2$.

Lemma 4: Under the probabilistic as well as the bounded overflow partial sharing models, for any $(x_1, x_2) \in \mathcal{X}^o$, *there exists* $\theta > 0$ *such that*

$$\nabla B_i(x_1, x_2) \cdot (1, \theta) < 0 \quad \forall i \in \{1, 2\}.$$

Lemma 4 implies that at any sharing configuration $x \in \mathcal{X}^o$, it is possible to strictly improve the blocking probability of both providers by increasing *both* components of x (in the direction θ).[4]

Below, we first use Lemma 4 to prove Theorem 2, and then prove Lemma 4.

Proof of Theorem 2: We provide a unified proof of Theorem 2 for both partial sharing models. Invoking Lemma 4 at the configuration $(0, 0)$, we conclude that the set of QoS-stable configurations is non-empty. For $i \in \{1, 2\}$, define

$$\mathcal{B}_i(x_1, x_2) := \max(0, E(N_i, a_i) - B_i(x_1, x_2)).$$

Consider the optimization $\max_{x \in \mathcal{X}} \mathcal{B}_1(x_1, x_2)\mathcal{B}_2(x_1, x_2)$. Since this is the maximization of a continuous function over a compact domain, a maximizer $x^* \in \mathcal{X}$ exists. It is easy to see that x^* is Pareto-efficient, implying that the set of Pareto-efficient configurations is non-empty. Finally, Lemma 4 implies that no Pareto-stable configuration lies in $\mathcal{X}^o$, implying that any Pareto-efficient configuration lies in $\mathcal{X} \setminus \mathcal{X}^o$. This completes the proof.

Proof of Lemma 4: We provide a unified proof of Lemma 4 for both partial sharing models. $\nabla B_1(x_1, x_2) \cdot (1, \theta) < 0$ is equivalent to

$$\theta > \frac{\frac{\partial B_1(x_1, x_2)}{\partial x_1}}{-\frac{\partial B_1(x_1, x_2)}{\partial x_2}} =: \underline{\theta}.$$

Similarly, $\nabla B_2(x_1, x_2) \cdot (1, \theta) < 0$ is equivalent to

$$\theta < \frac{-\frac{\partial B_2(x_1, x_2)}{\partial x_1}}{\frac{\partial B_2(x_1, x_2)}{\partial x_2}} =: \bar{\theta}.$$

We therefore have to prove that $\underline{\theta} < \bar{\theta}$, which is equivalent to

$$\frac{\partial B_1(x_1, x_2)}{\partial x_1}\frac{\partial B_2(x_1, x_2)}{\partial x_2} < \left(-\frac{\partial B_1(x_1, x_2)}{\partial x_2}\right)\left(-\frac{\partial B_2(x_1, x_2)}{\partial x_1}\right). \quad (1)$$

Since the blocking probabilities depend on λ_i and μ_i only through a_i, we consider two fictitious providers P_i' ($i \in \{1, 2\}$) with $\mu_1' = \mu_2' = 1$ and $\lambda_i' = \lambda_i/\mu_i$ such that $B_i' \equiv B_i$. For the providers P_i', we invoke Theorem 1, to deduce that $B_{ov}(x_1, x_2)$ is a strictly decreasing function of x_1 and x_2. This means

$$\lambda_1' \frac{\partial B_1(x_1, x_2)}{\partial x_1} < -\lambda_2' \frac{\partial B_2(x_1, x_2)}{\partial x_1} \quad (2)$$

$$\lambda_2' \frac{\partial B_2(x_1, x_2)}{\partial x_2} < -\lambda_1' \frac{\partial B_1(x_1, x_2)}{\partial x_2} \quad (3)$$

Noting that terms on both sides of (2) and (3) are positive, we can multiply the two inequalities to obtain (1).

It is important to note that even though Statement 3 of Theorem 1 assumes that $\mu_1 = \mu_2$, *the present proof does not.*

While Theorem 2 states that the (non-empty) set of Pareto-efficient configurations lies on the boundary of the space of partial sharing configurations (specifically, in the set $\mathcal{X} \setminus \mathcal{X}^o$), it does not provide a precise characterization of this set. Interestingly, such a precise characterization is possible, which is the goal of the following lemma.

Lemma 5: Under the probabilistic as well as the bounded overflow partial sharing models, the set $\hat{\mathcal{P}}$ *of Pareto-efficient sharing configurations is characterized as follows.*

1) *If* $E(N_1 + N_2, a_1 + a_2) < E(N_i, a_i)\ \forall\ i$, *then there exist uniquely defined constants* $\hat{x}_1$ *and* $\hat{x}_2$, *such that for* $i = 1, 2$, $\hat{x}_i \in (0, 1)$,

$$B_1(1, \hat{x}_2) = E(N_1, a_1),$$
$$B_2(\hat{x}_1, 1) = E(N_2, a_2).$$

[4]It is not hard to see that the blocking probabilities under the probabilistic sharing model (characterized in Lemma 1) are continuously differentiable over $\mathcal{X}$. For the bounded overflow model, the blocking probabilities (characterized in Lemma 2) are continuous over $[0, N_1] \times [0, N_2]$ and differentiable for $k_1, k_2 \notin \mathbb{Z}_+$. If k_i is an integer, then the partial left and right derivatives with respect to k_i exist. Thus, for the bounded overflow model, the gradients in the statement of Lemma 4 are understood to be composed of the right derivative with respect to x_i when $x_i N_i$ is an integer.

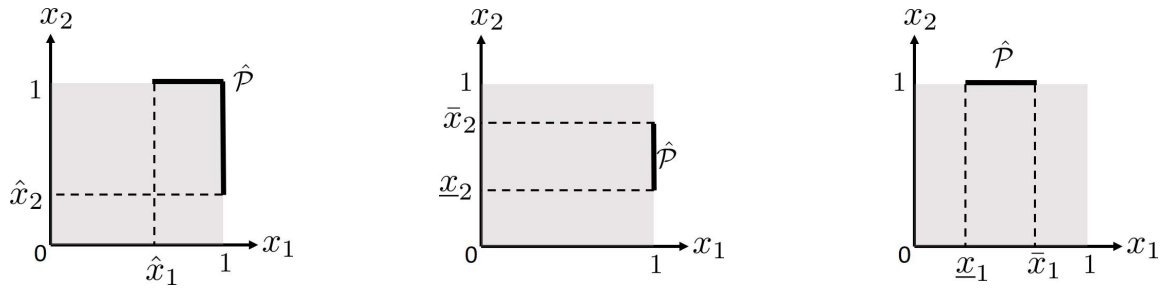


Fig. 1. The set $\hat{\mathcal{P}}$ of Pareto-efficient partial sharing configurations. (a) Case 1: Both providers are strictly better off under full pooling. (b) Case 2: Only Provider 1 is strictly better off under full pooling. (c) Case 3: Only Provider 2 is strictly better off under full pooling.

In this case,

$$\hat{\mathcal{P}} = \{(x, 1) : \ x \in (\hat{x}_1, 1]\} \cup \{(1, x) : \ x \in (\hat{x}_2, 1]\}.$$

2) *If* $E(N_2, a_2) \leq E(N_1 + N_2, a_1 + a_2) < E(N_1, a_1)$, *then there exist uniquely defined constants* $\underline{x}_2$ *and* $\bar{x}_2$ *satisfying* $0 < \underline{x}_2 < \bar{x}_2 \leq 1$ *such that*

$$B_1(1, \underline{x}_2) = E(N_1, a_1),$$
$$B_2(1, \bar{x}_2) = E(N_2, a_2).$$

In this case,

$$\hat{\mathcal{P}} = \{(1, x) : \ x \in (\underline{x}_2, \bar{x}_2)\}.$$

3) *If* $E(N_1, a_1) \leq E(N_1 + N_2, a_1 + a_2) < E(N_2, a_2)$, *then there exist uniquely defined constants* $\underline{x}_1$ *and* $\bar{x}_1$ *satisfying* $0 < \underline{x}_1 < \bar{x}_1 \leq 1$ *such that*

$$B_2(\underline{x}_1, 1) = E(N_2, a_2),$$
$$B_1(\bar{x}_1, 1) = E(N_1, a_1).$$

In this case,

$$\hat{\mathcal{P}} = \{(x, 1) : \ x \in (\underline{x}_1, \bar{x}_1)\}.$$

Figure 1 provides a pictorial representation of the set of Pareto-efficient partial sharing configurations under the three cases considered in Lemma 5. Note that Case 1 corresponds to settings where full pooling is beneficial to both providers. Cases 2 and 3 cover the more asymmetric settings, where exactly one provider (the more congested one) stands to benefit from full pooling. Lemma 5 states that in such cases, the more congested provider pools all of its servers under any Pareto-efficient sharing configuration. Intuitively, this is because the asymmetry in the value of servers pooled by each provider to the other. Indeed, servers pooled by the more congested provider add less value, since those servers are available for overflow calls of the less congested provider less often. As a result, mutually beneficial sharing configurations have the more congested provider pool more servers than the less congested provider.

The proof of Lemma 5 is provided in Appendix B.

The computation of the Pareto-frontier can be broken down into two steps. First we determine which provider pools fully, for which we need to compare $E(N_i, a_i)$ and $E(N_1 + N_2, a_1 + a_2)$. This can be accomplished in $O\left((N_1 + N_2)^2\right)$ steps (i.e., computations of the product form expression). Next we need to determine the end-points of the set. This can be accomplished to an accuracy of ϵ via binary search in $O\left((N_1 + N_2)^2 \log(1/\epsilon)\right)$ steps.

IV. Economics of Partial Sharing

The set $\hat{\mathcal{P}}$ of Pareto-efficient configurations characterized in Section III contains all possible sharing configurations which are minimal for the partial order induced by the usual relation "$\leq$" applied component-wise on the vectors of possible blocking probabilities. In other words, for every QoS-stable configuration outside of this set, there exists a configuration within $\hat{\mathcal{P}}$ that improves the blocking probability of at least one provider without worsening the blocking probability for the other. Unfortunately, the configurations within the Pareto set are not comparable under this component-wise relation. If we take any two configurations inside this set, then a configuration that is better for one of the providers will be worse for the other provider. Thus, rational providers who want to minimize their blocking probability will agree that it is beneficial for both of them to choose a configuration inside the Pareto set rather than one outside of this set, but may disagree on the choice of the configuration within the Pareto set.

It is then the natural to ask: Which configuration within the Pareto set should the two providers choose? Of course, in addition to the choices within the Pareto set, they could also choose not to share. This question, in a more general setting, has been investigated inside the framework of *bargaining theory*. In a typical two-player bargaining problem, two players have to agree upon one option amongst several. If both agree upon the option, then each player gets a utility corresponding to this option. On the other hand, if they fail to arrive at a consensus, then they get a utility corresponding to that of a *disagreement point*. In our setting, the two players are the two providers who have to agree on a partial sharing configuration (x_1, x_2) to implement. In case they are unable to reach such an agreement, there would be no resource pooling, i.e., the disagreement point is the configuration $(0, 0)$.

Our aim in this section is to present some of the most common solution concepts from bargaining theory and apply them to the partial resource sharing problem under consideration. We also present results of numerical experiments for different realistic network settings, highlighting the potential benefits of partial resource pooling in practice. Note that the discussion in this section applies to both the partial pooling models defined in Section II, although the numerical evaluations are restricted to the bounded overflow model due to space constraints.

A. Bargaining Solutions

The usual way to compute a solution of a bargaining problem is to first fix a set of axioms that a solution must satisfy. Axioms that appear often (though not necessarily

together) are Pareto optimality (PO), Symmetry (SYM), Scale Invariance (SI), Independence of Irrelevant Alternatives (IIA) and Monotonicity (MON). The interested reader is referred to [14] for a textbook treatment of axiomatic bargaining theory.

In addition to the axioms, some solution concepts rely on the convexity of the space of feasible utility pairs in order to guarantee uniqueness. In the present setting, the utility of a provider is a strictly decreasing function of its blocking probability. Due to space constraints, we restrict our attention to the linear case, i.e., the utility of P_i is taken to be $-B_i$, where B_i denotes its blocking probability. Numerical experiments show that this utility space is not convex. The usual method to overcome this drawback is to convexify the utility space by considering its convex hull. For our problem, this could lead to a solution of the form (as an example): configuration (k_1, k_2) with probability p and (k_1', k_2') with probability $(1-p)$. While on an abstract level, a solution in an extended space is acceptable, in practice its implementation is not straightforward because it needs to address questions like the following. Should the probability p be interpreted as a fraction of time during which (k_1, k_2) is implemented? If so, at what time-scale should the changes in configuration occur?

Another method of getting around convexity is to modify the set of axioms and show that some variation of the solutions concepts for the convex case satisfy them (see [15] and references therein). These however require some other assumptions on the utility set such as comprehensiveness[5] which is again difficult to verify in our setting.

We now apply four bargaining solutions from the literature to our partial pooling model. These are the Nash, Kalai-Smorodinsky, egalitarian and utilitarian bargaining solutions. The main result in this section shows the uniqueness of the Kalai-Smorodinsky and the egalitarian solutions without calling upon the standard arguments of convexity or comprehensiveness. The proof is based upon monotonicity properties highlighted in Section II.

For the bargaining solutions in this section, we assume that the utility of each provider is the negative of its blocking probability. In some situations, it may be more meaningful to take the negative logarithm of the blocking probability as the utility of a provider. We give the logarithmic variants of the Nash, Kalai-Smorodinsky, and the egalitarian solutions in Appendix E of [13].

Nash Bargaining Solution: The first concept we present was proposed by Nash in the seminal paper [16].

Definition 3: A partial sharing configuration (x_1^*, x_2^*) *is* Nash bargaining solution (NBS)*, if the partial sharing configuration satisfies the following condition*:

$$(x_1^*, x_2^*) = \arg\max_{x\in[0,1]^2} \left[B_1(0,0) - B_1(x_1,x_2)\right]_+ \times \left[B_2(0,0) - B_2(x_1,x_2)\right]_+ .$$

Here $[z]_+ := max(z, 0)$ denotes the positive part of x. At the NBS, the players are maximizing the product of the individual utilities relative to the disagreement point.[6] Clearly, any maximizer would lie in the set of $\hat{\mathcal{P}}$. However, the drawback of the NBS for our problem is that the utility space is not convex (observed in numerical experiments) which implies that the NBS may not be unique.

Kalai-Smorodinsky Bargaining Solution: One of criticisms of the NBS is the axiom of IIA which may not hold in practice. In [17], Kalai and Smorodinsky replaced IIA with MON and obtained the following solution concept.

Definition 4: A partial sharing configuration (x_1^*, x_2^*) *is a* Kalai-Smorodinsky bargaining solution (KSBS)*, if* $(x_1^*, x_2^*) \in \hat{\mathcal{P}}$ *and satisfies*

$$\frac{B_1(0,0) - B_1(x_1,x_2)}{B_2(0,0) - B_2(x_1,x_2)} = \frac{B_1(0,0) - \min\limits_{y\in[0,1]^2} B_1(y_1,y_2)}{B_2(0,0) - \min\limits_{y\in[0,1]^2} B_2(y_1,y_2)}.$$

At a KSBS solution the ratio of relative utilities of the providers is equal to the ratio of their maximal relative utilities. For our problem, the following result guarantees uniqueness of the KSBS, which makes it potentially more attractive than the NBS.[7]

Theorem 3: For the bounded overflow sharing model, the KSBS is unique.

Proof of Theorem 3: Define the following functions.

$$f(x_1,x_2) := \frac{B_1(0,0) - B_1(x_1,x_2)}{B_2(0,0) - B_2(x_1,x_2)}$$

$$KS := \frac{B_1(0,0) - \min\limits_{y\in[0,1]^2} B_1(y_1,y_2)}{B_2(0,0) - \min\limits_{y\in[0,1]^2} B_2(y_1,y_2)}$$

From the Statements 1 and 2 of Theorem 1, we get

$$\min_{y\in[0,1]^2} B_1(y_1,y_2) = B_1(0,1),$$
$$\min_{y\in[0,1]^2} B_2(y_1,y_2) = B_2(1,0).$$

i.e., each provider gets the maximum benefit when it pools none of its servers and the other provider pools all of its servers.

It is easy to see that $0 < KS < \infty$. Consider the three cases for $\hat{\mathcal{P}}$ from Lemma 5.

Case 1 ($E(N_1 + N_2, a_1 + a_2) < E(N_i, a_i)\ \forall\ i$): Sweeping the (topologically one-dimensional) Pareto-frontier clockwise from $(\hat{x}_1, N_2)$ to $(N_1, \hat{x}_2)$, it is easy to see that f is strictly decreasing and continuous, with

$$\lim_{x_1\downarrow\hat{x}_1} f(x_1,1) = \infty,$$
$$\lim_{x_2\downarrow\hat{x}_2} f(1,x_2) = 0.$$

There is thus a unique point on the Pareto-frontier that satisfies the KSBS condition.

Case 2 ($E(N_2, a_2) \le E(N_1 + N_2, a_1 + a_2) < E(N_1, a_1)$): As before, sweeping the Pareto-frontier clockwise from $(1, \bar{x}_2)$ to $(1, \underline{x}_2)$, it is easy to see that f is strictly decreasing and continuous, with

$$\lim_{x_2\uparrow\bar{x}_2} f(1,x_2) = \infty,$$
$$\lim_{x_2\downarrow\underline{x}_2} f(1,x_2) = 0.$$

[5]Comprehensiveness says that for any vector in the utility set, all vectors that are weakly dominated by this vector and that weakly dominate the disagreement point are also in the utility set.

[6]Here, *relative* means upon subtracting the utilities at the disagreement point.

[7]As is apparent from the proof of Theorem 3, the KSBS can be computed via a binary search in $O\left((N_1+N_2)^2 \log(1/\epsilon)\right)$ steps, given an accuracy threshold ϵ.

There is thus a unique point on the Pareto-frontier that satisfies the KSBS condition.

Case 3 ($E(N_1, a_1) \leq E(N_1 + N_2, a_1 + a_2) < E(N_2, a_2)$): The argument here is analogous to that for the above cases.

Egalitarian Solution: The next solution concept we present was also proposed by Kalai [18]. It satisfies PO, SYM, IIA, and MON but violates SI. It captures the sharing configuration in which the gains relative to the disagreement solution for both the providers is the same.

Definition 5: A partial sharing configuration (x_1^*, x_2^*) is an egalitarian solution (ES)*, if $(x_1^*, x_2^*) \in \hat{\mathcal{P}}$ and satisfies*

$$B_1(0,0) - B_1(x_1, x_2) = B_2(0,0) - B_2(x_1, x_2).$$

Under an ES, the providers will see the same amount of improvement in their blocking probabilities relative to the no-sharing option. The following result shows that the ES is unique. Its proof follows similar lines as the proof of Theorem 3.[8]

Lemma 6: For the bounded overflow sharing model, the ES is unique.

Proof of Lemma 6: The argument in the proof of Theorem 3 is applicable without change here, except that the constant KS is replaced by 1.

An interesting property of the ES is that if the standalone blocking probabilities of the two providers are identical, then the ES corresponds to complete pooling.

Lemma 7: If $E(N_1, a_1) = E(N_2, a_2)$, then the ES lies at $(1, 1)$.

Proof of Lemma 7: We invoke the following well known property of the Erlang-B formula.

$$E(N_1+N_2, a_1+a_2) < \frac{a_1}{a_1+a_2} E(N_1, a_1) + \frac{a_2}{a_1+a_2} E(N_2, a_2).$$

If $E(N_1, a_1) = E(N_2, a_2)$, it follows then that

$$E(N_1 + N_2, a_1 + a_2) < E(N_1, a_1) = E(N_2, a_2),$$

implying that the set $\hat{\mathcal{P}}$ of Pareto-efficient configurations includes $(1, 1)$ (see Lemma 5).

Further, is $E(N_1, a_1) = E(N_2, a_2)$, then the ES clearly satisfies $B_1(x_1, x_2) = B_2(x_1, x_2)$. However, from the monotonicity properties of the blocking probabilities, $(1, 1)$ is the only point in $\hat{\mathcal{P}}$ that satisfies this property.

Utilitarian Solution: The last solution concept is that of utilitarian bargaining solution (see, e.g., [19]). It minimizes the blocking probability of the customers as a whole without distinguishing them according the provider to which they subscribe. It captures the greatest good to the system. The advantage is that it is a concept that is easy for customers to identify with. On the other hand, the axioms of SI and MON are violated. Nonetheless, the violation of SI does not seem to be problematic when the utilities are blocking probabilities. Indeed, there is a unique natural scale on which the blocking probability satisfies the axioms that define a probability measure.

Definition 6: A partial sharing configuration (x_1^*, x_2^*) is a utilitarian bargaining solution *(US) if it satisfies*

$$\arg\min_{k \in \boldsymbol{C}(\hat{\mathcal{P}})} \frac{\lambda_1}{\lambda_1 + \lambda_2} B_1(x_1 x_2) + \frac{\lambda_2}{\lambda_1 + \lambda_2} B_2(x_1, x_2).$$

[8]The computational complexity of the egalitarian solution is the same as the KSBS.

Here, $\boldsymbol{C}(\hat{\mathcal{P}})$ denotes the closure of $\hat{\mathcal{P}}$. We relax the above minimization to be over $\boldsymbol{C}(\hat{\mathcal{P}})$ instead of over the open set $\hat{\mathcal{P}}$ because in some cases, it turns out that the solution lies on the boundary. Assuming that the average call holding time for both providers is identical, the utilitarian solution is unique and can be characterized precisely.[9]

Lemma 8: If $\mu_1 = \mu_2$, under the bounded overflow model, the US is characterized as follows.[10]

1) *If $E(N_1 + N_2, a_1 + a_2) < E(N_i, a_i)\ \forall\ i$, then the US is $(1, 1)$*
2) *If $E(N_2, a_2) \leq E(N_1 + N_2, a_1 + a_2) < E(N_1, a_1)$, then the US is $(1, \bar{x}_2)$*
3) *If $E(N_1, a_1) \leq E(N_1 + N_2, a_1 + a_2) < E(N_2, a_2)$, then the US is $(\bar{x}_1, 1)$*

We omit the proof of Lemma 8, since it is direct consequence of Statement 3 of Theorem 1. Another quick observation is that when the standalone blocking probabilities are matched, the utilitarian solution, like the egalitarian solution, corresponds to full pooling.

Corollary 1: If $E(N_1, a_1) = E(N_2, a_2)$, then the US lies at $(1, 1)$.

Proof of Corollary 1: As was argued in the proof of Lemma 7, if $E(N_1, a_1) = E(N_2, a_2)$, then

$$E(N_1 + N_2, a_1 + a_2) < E(N_1, a_1) = E(N_2, a_2).$$

The statement of the corollary now follows from Lemma 8.

While the utilitarian solution is the most efficient, in that is minimizes the overall blocking probability, it may not be fair. Indeed, under Cases 2 and 3 of Lemma 8 above, one of the providers (the less congested provider) sees no reduction in its blocking probability relative to the disagreement point.

B. Numerical Examples

In this section, we present numerical results illustrating the various bargaining solutions under realistic system settings. The goal of this section is two-fold: to demonstrate the benefits of partial resource pooling to the two providers, and to illustrate differences between the different bargaining solutions. Due to space constraints, we are only able to consider two network settings. Also, we restrict our attention in this section to the bounded overflow sharing model; we represent the bargaining solution as (k_1^*, k_2^*), where $k_i^* = N_i x_i^*$.

The system parameters for this section are chosen to capture realistic scenarios in cellular networks. The number of simultaneous calls that can be handled by a typical cell tower range from 30 to 200, depending on spectrum allocation and the technology in use [1], [20]. Moreover, regulators worldwide mandate that call drop rates should be at most 2%, whereas empirical studies have recorded drop rates exceeding 5% in dense urban environments [21], [22].

First we consider a scenario where the two providers have the same number of servers, but differ with respect to their standalone blocking probabilities. Specifically, we set $N_1 = N_2 = 100$, with $E(N_1, a_1) = 0.06$ (6%), $E(N_1, a_1) = 0.01$ (1%). Clearly P_1 is the more congested provider. The different bargaining solutions for this scenario are summarized

[9]It is clear from Lemma 8 that the computational complexity of the utalitarian solution is the same as that of computing the end-points of the Pareto frontier, which is discussed in Section III.

[10]We use the notation from Lemma 5.

TABLE I

DIFFERENT BARGAINING SOLUTIONS FOR THE CASE $N_1 = N_2 = 100$, THE STANDALONE BLOCKING PROBABILITIES OF P_1 AND P_2 BEING 6% AND 1%, RESPECTIVELY

Bargaining solution	k_1^*	k_2^*	B_1	B_2
US	100	13.1	1. 73%	1%
KSBS	100	6	3.39%	0.63 %
NBS	100	5.5	3.6%	0.6 %
ES	100	1.35	5.36%	0.36%

TABLE II

DIFFERENT BARGAINING SOLUTIONS FOR THE CASE $N_1 = 200$, $N_2 = 50$, BOTH PROVIDERS HAVING A STANDALONE BLOCKING PROBABILITY OF 5%

Bargaining solution	k_1^*	k_2^*	B_1	B_2
US	200	50	3.42%	3.42%
ES	200	50	3.33%	3.33%
NBS	200	9.5	3.36%	3.19%
KSBS	200	8	3.56%	2.99%

in Table I. As expected, the more congested provider P_1 pools all its servers under all bargaining solutions. Moreover, the 'efficient' utilitarian solution is the most beneficial for P_1, while not providing any benefit to P_2. At the other extreme, ES is the most pessimal, since it enforces the same reduction in blocking probability, even though the scope for reduction is much less for P_2. KSBS and NBS result in intermediate contributions by P_2, and result in a substantial benefits for both P_1 and P_2; indeed, these configurations result in a roughly 40% reduction in the blocking probability of each provider.

Next, we consider a scenario where the two providers differ in size, but are matched with respect to standalone blocking probability. Specifically, we set $N_1 = 200$, $N_2 = 50$, $E(N_1, a_1) = E(N_2, a_2) = 0.05$. The results are summarized in Table II. As expected, the US as well as the ES correspond to complete pooling (see Lemma 7 and Corollary 1); this results in both providers seeing a blocking probability of 3.33%. On the other hand, the NBS as well as the KSBS, the smaller provider (P_2) pools fewer servers. As a result, the smaller provider achieves an even lower blocking probability under KSBS/NBS, at the expense of a higher blocking probability for the larger provider (compared to the full pooling under US/ES). As before, it is important to note that partial resource pooling offers the possibility of substantially lower blocking probability for both providers.

V. LARGE SYSTEM LIMITS: SQUARE ROOT SCALING

The computational complexity of the exact steady-state blocking probability increases as the number of circuits becomes large [9]. As a result, approximations can turn out to be helpful for their tractability as well as their ability to provide insights into the complex dependencies between the blocking probabilities and the system parameters. The goal of this section is to obtain large system approximations for the blocking probabilities under the bounded overflow partial pooling model.[11]

Large system approximations have always been an integral part of the literature on queuing theory. Depending upon the parameters of systems, these limits can take different forms such as mean-field [23], Quality and Efficiency Driven [24], or Non-degenerate Slowdown [25] limits.

A. QED Scaling Regime

For our resource sharing model with blocking, the most relevant limit is the quality-efficiency-driven (QED) regime (a.k.a. "square-root staffing" regime, Halfin-Whitt regime). While it is now commonly known under these names, it had already been investigated by Erlang himself[12] and Jagerman as well [27]. The traditional QED regime applies to a system with a single provider, and is defined as follows. Let N be the number of circuits with the provider and a be the offered load. We say that $f(t) \sim g(t)$ as $t \to \infty$ if $\lim_{t\to\infty} \frac{f(t)}{g(t)} = 1$.

Lemma 9 [27]: Let $a = N + \beta\sqrt{N} + o(\sqrt{N})$. Then,

$$E(N, a) \sim \frac{1}{\sqrt{N}} \frac{\phi(\beta)}{(1 - \Phi(\beta))} \quad \text{as } N \to \infty.$$

Here, $\phi(\cdot)$ and $\Phi(\cdot)$ denote, respectively, the probability density function and the cumulative distribution function, corresponding to the standard Gaussian distribution. Note that under the QED regime, the margin between the offered load and the number of servers is of the order of the square root of the number of servers. In many settings, the QED regime is known to be the right balance between quality (i.e., QoS) and efficiency (i.e., server provisioning costs); see, for example, [24], [28]. For the M/M/N/N loss system, Lemma 9 states that the steady state blocking probability decays as $\Theta(1/\sqrt{N})$ as $N \to \infty$.

We define the QED scaling regime for our model with two providers as follows. For fixed $\alpha_i > 0$ and $\beta_i \in \mathbb{R}$, let

$$N_i = \alpha_i N, \tag{4}$$

$$a_i = N_i + \beta_i \sqrt{N_i} + o(\sqrt{N_i}). \tag{5}$$

Here, N is the scaling parameter that is common to both providers. (4) states that the number of servers of each provider grows proportionately with the scaling parameter. (5) states that the offered load corresponding to each provider scales as per the QED (square-root staffing) rule.

Before deriving the blocking probabilities for the different partial sharing configurations, we first look at two special cases for which these probabilities can be derived directly from Lemma 9. With no resource pooling, both the providers are decoupled, and for large N, the steady state blocking probability of Provider i can be computed using Lemma 9 to be

$$B_i \sim \frac{1}{\sqrt{N_i}} \frac{\phi(\beta_i)}{(1 - \Phi(\beta_i))}.$$

The second special case is that of full resource pooling. Here, the system acts as a single provider with $(N_1 + N_2)$ servers/circuits and offered load of $(a_1 + a_2)$. By simple calculations we can see the system under full pooling also

[11] A parallel development for the probabilistic sharing model is possible, which we omit due to space constraints.

[12] See the paper "On the rational determination of the number of circuits" in [26].

satisfies the square root scaling set up. So, the steady state blocking probability for both the providers is given as

$$B_{full} \sim \frac{1}{\sqrt{N(\alpha_1+\alpha_2)}} \frac{\phi\left(\frac{\beta_1\sqrt{\alpha_1}+\beta_2\sqrt{\alpha_2}}{\sqrt{\alpha_1+\alpha_2}}\right)}{1-\Phi\left(\frac{\beta_1\sqrt{\alpha_1}+\beta_2\sqrt{\alpha_2}}{\sqrt{\alpha_1+\alpha_2}}\right)}.$$

Now, we present the square-root scaling set up for partial sharing configurations. For $\gamma_i \geq 0$, we scale the sharing parameters as

$$k_i = \gamma_i\sqrt{N_i} + o(\sqrt{N_i}). \tag{6}$$

Note that the number of pooled servers for P_i is scaled in proportion to $\sqrt{N_i}$. It turns out that for the system scaling defined by (4)–(5), this is the only meaningful manner of scaling the partial sharing parameters. Indeed, if $k_i = o(\sqrt{N_i})$, then the large system limits correspond to P_i pooling no servers, and if $k_i = \omega(\sqrt{N_i})$, then the large system limits correspond to P_i pooling all its servers. Intuitively, this is because on the diffusion scale defined by (4)–(5), the number of overflow calls as well as the number of free servers of each provider evolve (in time) on the $\sqrt{N}$ scale.

To summarize, the QED regime we consider is defined by (4)–(6). Our main result in this section gives the relationship between the asymptotic blocking probability for each provider and the various parameters of the system, namely, the sharing parameters (γ_1, γ_2), the square-root staffing margins (β_1, β_2), and the relative sizes of the two providers (α_1, α_2).

B. Blocking Probability Asymptotics

Having defined our QED scaling regime, we now derive large system asymptotics of the blocking probabilities. Our results are summarized in the following theorem.

Theorem 4: Under the bounded overflow sharing model, for the scaling regime defined by (4)–(6)*, the steady state blocking probability of Provider i for large N is given as*

$$B_i(\gamma_1,\gamma_2) \sim \frac{1}{\sqrt{N}}\frac{\tilde{A}_i}{\tilde{G}},$$

where

$$\begin{aligned}
\tilde{A}_i &= \frac{\phi\left(\gamma_{-i}\sqrt{\frac{\alpha_{-i}}{\alpha_i}} - \beta_i\right)}{\sqrt{\alpha_i}}\Phi\left(-\gamma_{-i}-\beta_{-i}\right) \\
&\quad + \frac{1}{\sqrt{\alpha_1\alpha_2}}\int_{-\gamma_1\sqrt{\alpha_1}}^{\gamma_2\sqrt{\alpha_2}} \phi\left(\frac{x}{\sqrt{\alpha_1}}-\beta_1\right)\phi\left(\frac{-x}{\sqrt{\alpha_2}}-\beta_2\right)dx, \\
\tilde{G} &= \Phi\left(\gamma_1\sqrt{\frac{\alpha_1}{\alpha_2}}-\beta_2\right)\Phi\left(-\gamma_1-\beta_1\right) \\
&\quad + \frac{1}{\sqrt{\alpha_1}}\int_{-\gamma_1\sqrt{\alpha_2}}^{\gamma_2\sqrt{\alpha_1}} \phi\left(\frac{x}{\sqrt{\alpha_1}}-\beta_1\right)\Phi\left(\frac{-x}{\sqrt{\alpha_2}}-\beta_2\right)dx.
\end{aligned}$$

Even though the expressions for $\tilde{A}_i$ and G in the statement of Theorem 4 look complicated, they have a simple geometric interpretation. To see this, define the following sets in $\mathbb{R}^2$.

$$\begin{aligned}
\tilde{M} &:= \left\{(x_1,x_2): \begin{smallmatrix} x_1\leq\gamma_2\sqrt{\alpha_2} \\ x_2\leq\gamma_1\sqrt{\alpha_1} \\ x_1+x_2\leq 0\end{smallmatrix}\right\}, \\
\tilde{R} &:= \left\{(x_1,x_2): \begin{smallmatrix} x_1\leq\gamma_2\sqrt{\alpha_2} \\ x_2\leq\gamma_1\sqrt{\alpha_1} \\ x_1+x_2= 0\end{smallmatrix}\right\}, \\
\tilde{C}_1 &:= \{(x_1,\gamma_1\sqrt{\alpha_1}):\ x_1\leq-\gamma_1\sqrt{\alpha_1})\}, \\
\tilde{C}_2 &:= \{(\gamma_2\sqrt{\alpha_2},x_2):\ x_2\leq-\gamma_2\sqrt{\alpha_2})\}.
\end{aligned}$$

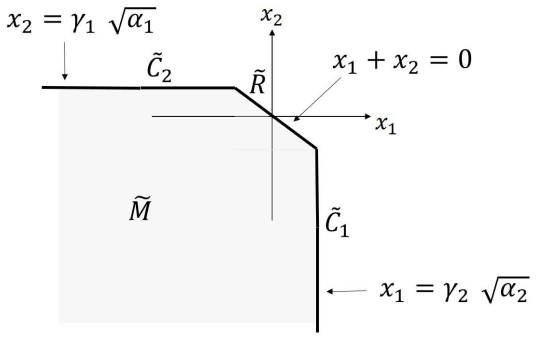


Fig. 2. Geometric interpretation of blocking probability asymptotics under QED.

These sets are depicted in Figure 2. Note that $\tilde{M}$ is the shaded pentagonal region, and $\tilde{R}$, $\tilde{C}_1$, and $\tilde{C}_2$, represent the diagonal, right, and upper boundaries of $\tilde{M}$, respectively. Now, define independent Gaussian random variables Z_1 and Z_2, such that Z_i has mean $\beta_i\sqrt{\alpha_i}$, and variance α_i. Let $f_{Z_i}(\cdot)$ denote the probability density function corresponding to Z_i. With this notation, it is not hard to show that

$$\begin{aligned}
\tilde{A}_i &= \int_{\tilde{R}} f_{Z_1}(x_1)f_{Z_2}(x_2) + \int_{\tilde{C}_i} f_{Z_1}(x_1)f_{Z_2}(x_2), \\
\tilde{G} &= \iint_{\tilde{M}} f_{Z_1}(x_1)f_{Z_2}(x_2)
\end{aligned} \tag{7}$$

This means that $\tilde{A}_i$ is the line integral of the joint density function of Z_1 and Z_2 over $\tilde{R}\cup\tilde{C}_i$, and $\tilde{G}$ is the integral of the same joint density function over the region $\tilde{M}$ (in other words, $\tilde{G}$ is the probability that the random vector (Z_1, Z_2) takes a value in $\tilde{M}$).

Theorem 4 yields a computationally tractable approximation for the blocking probabilities under the BO partial sharing model, which is asymptotically accurate under the QED regime. In particular, note that the computational complexity of the approximation is invariant to the system size, making it particularly attractive when the number of servers is large. In the remainder of this section, we evaluate the accuracy of the large system approximation under realistic network settings. The proof of Theorem 4 can be found in Appendix C of [13].

C. Accuracy of Large System Approximation

Due to space constraints, we do not directly illustrate the accuracy of the large system approximation here; the interested reader is referred to [13] for these calculations. Instead, we focus here on the accuracy of the bargaining solutions computed using our large system approximation.

Specifically, for the same network settings used in Section IV-B, we compute the various bargaining solutions

TABLE III

BARGAINING SOLUTIONS COMPUTED USING THE LARGE SYSTEM APPROXIMATION: $N_1 = N_2 = 100$; STANDALONE BLOCKING PROBABILITIES OF P_1 AND P_2 ARE 6% AND 1%, RESPECTIVELY

Bargaining solution	k_1^*	k_2^*	B_1	B_2
US	100	11	2.22%	1%
KSBS	100	5.7	3.71%	0.68 %
NBS	100	5.2	3.9%	0.65 %
ES	100	1.7	5.46%	0.44%

TABLE IV

BARGAINING SOLUTIONS COMPUTED USING THE LARGE SYSTEM APPROXIMATION: $N_1 = 200$, $N_2 = 50$; BOTH PROVIDERS HAVING A STANDALONE BLOCKING PROBABILITY OF 5%

Bargaining solution	k_1^*	k_2^*	B_1	B_2
US	200	50	3.42%	3.42%
ES	200	50	3.42%	3.42%
NBS	200	10.3	3.69%	2.93%
KSBS	200	8.8	3.81%	2.74%

using the large system approximation, and compare these against the same solutions computed using the exact (but more computationally intensive) blocking probability formulae. The results are summarized in Tables III and IV. Comparing these results against the corresponding calculations in Tables I and II respectively, we see that the large system approximations provide good approximations of the bargaining solutions. But more importantly, note that the relative ordering between the different bargaining solutions is retained by the large system approximations. This indicates that our approximations are useful for understanding bargaining outcomes in practical loss systems.

VI. DISCUSSION

We conclude with a discussion on some analogies from circuit multiplexed networks and possible generalizations.

Circuit Multiplexed Network Analogs: When the k_i are integers, under the bounded overflow model, the state space and the stationary distribution will be the same as a circuit multiplexed network of with two routes and three links (one link being common to both routes. With this representation the reduced load approximation method of, e.g., [29] may also be used to calculate the blocking probabilities.

As was mentioned in Section II, there is a superficial similarity between the BO model and trunk reservation. Trunk reservation has been used in circuit multiplexed networks to give preference to direct route calls over alternate route calls. This is done by reserving the 'last k circuits' for direct route calls. This means that on a link with N circuits, alternate route calls are not admitted when the number of idle circuits is less than or equal to k. Exact models for trunk reservation are hard to analyze and asymptotic analyses, e.g., [30], are among analytical techniques that have been used to model trunk reservation.

Future Work: Several extensions of partial pooling to models in extant literature are possible. Partial sharing among multiple providers, and over Kelly networks are obvious extensions that we are considering at this time. It must be noted that simple bilateral sharing models are hard to define in these systems. A second extension would be to consider Erlang-C waiting systems. Here too, defining suitable, and analytically tractable, sharing models appears to be non-trivial.

APPENDIX A
PROOF OF THEOREM 1 FOR THE BOUNDED OVERFLOW SHARING MODEL

The following well known properties of the Erlang-B formula will be useful.

Lemma 10: *$E(N,a)$ is a strictly decreasing function of N. Moreover, $E(N,a) > 1 - \frac{N}{a}$.*

We also state the following lemma which will be invoked repeatedly in the proof.

Lemma 11: Under the bounded overflow sharing model, if $k_i \in \{1, 2, \cdots, N_i\}$ $\forall i$,

$$E(N_i + k_{-i}, a_i) < B_i^{(bo)}(k_1, k_2) < E(N_i - k_i, a_i) \qquad (i \in \{1, 2\}).$$

The proof is elementary and is omitted.

Proof of Statements 1 and 2: The steady state blocking probability of Provider 1 can be expressed as follows.

$$B_1^{(bo)}(k_1, k_2) = \frac{m_1 + u_1\{k_1\} + v_1\{k_2\}}{d + u\{k_1\} + v\{k_2\}}$$

Here,

$$m_1 = \sum_{i=N_1-\lfloor k_2 \rfloor}^{N_1+\lfloor k_2 \rfloor} \frac{a_1^i}{i!} \frac{a_2^{(N_1+N_2-i)}}{(N_1+N_2-i)!} + \frac{a_1^{(N_1+\lfloor k_2 \rfloor)}}{(N_1+\lfloor k_2 \rfloor)!} \times \sum_{j=0}^{N_2-\lfloor k_2 \rfloor - 1} \frac{a_2^j}{j!},$$

$$u_1 = \frac{a_1^{(N_1-\lceil k_1 \rceil)}}{(N_1-\lceil k_1 \rceil)!} \frac{a_2^{(N_2+\lceil k_1 \rceil)}}{(N_2+\lceil k_1 \rceil)!},$$

$$v_1 = \left(1 - \frac{N_1 + \lceil k_2 \rceil}{a_1}\right) \frac{a_1^{(N_1+\lceil k_2 \rceil)}}{(N_1+\lceil k_2 \rceil)!} \sum_{j=0}^{N_2-\lceil k_2 \rceil} \frac{a_2^j}{j!},$$

$$d = \sum_{\substack{(i,j):\ i \le N_1+\lfloor k_2 \rfloor \\ j \le N_2+\lfloor k_1 \rfloor \\ i+j \le N_1+N_2}} \frac{a_1^i}{i!} \frac{a_2^j}{j!},$$

$$u = \frac{a_2^{(N_2+\lceil k_1 \rceil)}}{(N_2+\lceil k_1 \rceil)!} \sum_{i=0}^{N_1-\lceil k_1 \rceil} \frac{a_1^i}{i!},$$

$$v = \frac{a_1^{(N_1+\lceil k_2 \rceil)}}{(N_1+\lceil k_2 \rceil)!} \sum_{j=0}^{N_2-\lceil k_2 \rceil} \frac{a_2^j}{j!}.$$

Since $B_1^{(bo)}(k_1, k_2)$ is continuous in its arguments, it suffices to show that for non-integer (k_1, k_2),

$$\frac{\partial B_1^{(bo)}(k_1,k_2)}{\partial k_2} < 0, \quad \frac{\partial B_1^{(bo)}(k_1,k_2)}{\partial k_1} > 0.$$

Accordingly, in the remainder of the proof, we make the assumption that $\{k_1\}, \{k_2\} \neq 0$.

We now prove that $\frac{\partial B_1^{(bo)}(k_1,k_2)}{\partial k_2} < 0$. An elementary calculation yields

$$\frac{\partial B_1^{(bo)}(k_1,k_2)}{\partial k_2} = \frac{(v_1\ d - v m_1) + (v_1 u - v u_1)\{k_1\}}{(G')^2}$$

It now suffices to show that each term in the numerator above is negative. To see that the first term is negative, note that

$$\begin{aligned} v_1 d - v m_1 &= v d \left(\frac{v_1}{v} - \frac{m_1}{d} \right) \\ &= v d \left(\left(1 - \frac{N_1 + \lceil k_2 \rceil}{a_1} \right) - B_1^{(bo)}(\lfloor k_1 \rfloor, \lfloor k_2 \rfloor) \right) \\ &< v d \left(E(N_1 + \lceil k_2 \rceil, a_1) - B_1^{(bo)}(\lfloor k_1 \rfloor, \lfloor k_2 \rfloor) \right) < 0. \end{aligned}$$

The first inequality above uses Lemma 10, and the second uses Lemma 11. To see that the second term is negative, note that

$$\begin{aligned} v_1 u - u_1 v &= v u \left(\frac{v_1}{v} - \frac{u_1}{u} \right) \\ &= v u \left(\left(1 - \frac{N_1 + \lceil k_2 \rceil}{a_1} \right) - E(N_1 - \lceil k_1 \rceil, a_1) \right) \\ &< v u \left(E(N_1 + \lceil k_2 \rceil, a_1) - E(N_1 - \lceil k_1 \rceil, a_1) \right) < 0. \end{aligned}$$

Both the above inequalities follow from Lemma 10. Therefore, we conclude that $\frac{\partial B_1^{(bo)}(k_1,k_2)}{\partial k_2} < 0$.

Next, we prove that $\frac{\partial B_1^{(bo)}(k_1,k_2)}{\partial k_1} > 0$. An elementary calculation yields

$$\frac{\partial B_1^{(bo)}(k_1,k_2)}{\partial k_1} = \frac{(u_1\ d - u m_1) + (u_1 v - u v_1)\{k_2\}}{(G')^2}$$

It suffices to argue that each of the terms in the numerator above is positive. To see that the first term is positive, note that

$$\begin{aligned} u_1 d - u m_1 &= u d \left(\frac{u_1}{u} - \frac{m_1}{d} \right) \\ &= u d \left(E(N_1 - \lceil k_1 \rceil, a_1) - B_1^{(bo)}(\lfloor k_1 \rfloor, \lfloor k_2 \rfloor) \right) > 0. \end{aligned}$$

The inequality above follows from Lemma 11. Since we have already proved that $(v_1 u - v u_1) < 0$, it follows that the second term is also positive. This proves that $\frac{\partial B_1^{(bo)}(k_1,k_2)}{\partial k_1} > 0$.

Proof of Statement 3: It suffices to prove that $B_{overall}^{(bo)}(k_1, k_2)$ is a strictly decreasing function of k_1. We first prove the monotonicity over integer-valued k_1 (Lemma 12) and then show that the monotonicity also extends to real-valued k_1.

Lemma 12: If $\mu_1 = \mu_2 = \mu$, then for $k_1 \in \{0, 1, \cdots, N_1 - 1\}$ and $k_2 \in [0, N_2]$, $B_{overall}^{(bo)}(k_1 + 1, k_2) < B_{overall}^{(bo)}(k_1, k_2)$.

Proof of Lemma 12: The proof is based on a sample path approach. We assume that call holding times for both providers are exponentially distributed with parameter μ (we are free to make this assumption given that the blocking probabilities are insensitive to the call holding time distributions).

We consider two systems – an 'old' system (O) with sharing configuration (k_1, k_2) and a 'new' system (N) with perturbed sharing configuration $(k_1 + 1, k_2)$. In what follows, we will couple the arrival processes and service durations across these systems in such a way that N system will serve at-least as many calls as the O system on any sample path.

At time zero, we start with both the N and the O system being empty. Let $n_i(t)$ denote the number of Provider i calls in the O system at time t, and $\tilde{n}_i(t)$ denote the number of Provider i calls in the N system at time t. The two systems see exactly the same call arrival process. Moreover, calls that are admitted into both systems have the same holding time (this ensures that such calls complete at the same time in both systems). Calls that are admitted into one of the systems but not into the other are categorized as follows.

- Type 1: A Provider 2 call that is admitted into the N system but not the O system.
- Type 2: A Provider 1 call that is admitted into the N system but not the O system.
- Type 3: A Provider 2 call that is admitted into the O system but not the N system.
- Type 4: A Provider 1 call that is admitted into the O system but not the N system.

Note that

- $n_1(t) - \tilde{n}_1(t) = \#$ of Type 4 calls (in O) at time $t - \#$ of Type 2 calls (in N) at time t
- $\tilde{n}_2(t) - n_2(t) = \#$ of Type 1 calls (in N) at time $t - \#$ of Type 3 calls (in O) at time t

We will now couple the service durations of Type j calls in such a way that at all times, the states of the N and O systems satisfy one of the following three relations:

- **R1**: $n_i(t) = \tilde{n}_i(t)$ for $i = 1, 2$
- **R2**: $n_1(t) = \tilde{n}_1(t)$, and $n_2(t) = \tilde{n}_2(t) - 1$
- **R3**: $n_1(t) = \tilde{n}_1(t) + 1$, and $n_2(t) = \tilde{n}_2(t) - 1$

Note that under all three relations,

$$\tilde{n}_1(t) + \tilde{n}_2(t) \geq n_1(t) + n_2(t), \tag{8}$$

with equality under R1 and R3, and a strict inequality under R2. Moreover,

$$n_1(t) \geq \tilde{n}_1(t), \quad \tilde{n}_2(t) \geq n_2(t). \tag{9}$$

The states satisfy R1 at time 0. Moreover, calls that get admitted into both systems do not alter the relation between the states, at times of arrival or departure. So we only need to focus on arrival/departure epochs of Type j calls, $j \in \{1, 2, 3, 4\}$. Our argument will proceed inductively in time.

Type 1 Arrival: Suppose that a Type 1 arrival (into N) occurs at time s. Since (8) holds at time s^-, we must have

$$n_1(s^-) + n_2(s^-) \leq \tilde{n}_1(s^-) + \tilde{n}_2(s^-) < N_1 + N_2. \tag{10}$$

This implies that at s^-, the states satisfy R1 with $n_2(s^-) = \tilde{n}_2(s^-) = N_2 + k_1$. In this case, the arrival would result in $\tilde{n}_2(s) = N_2 + k_1 + 1$, implying the states would satisfy R2 at time s. The holding time of the newly arrived call in N is taken to be an independent $\mathrm{Exp}(\mu)$ random variable.

Type 2 Arrival: Suppose that a Type 2 arrival (into N) occurs at time s. This implies (10) as before. It then follows that $n_1(s^-) \in \{N_1 + \lfloor k_2 \rfloor, N_1 + \lceil k_2 \rceil\}$, and $\tilde{n}_1(s^-) < n_1(s^-)$. This implies that the states satisfy R3 at s^-, and thus satisfy R2 at time s. In other words, at time s, we have

- $\#$ of Type 4 calls (in O) $= \#$ of Type 2 calls (in N)
- $\#$ of Type 1 calls (in N) $= \#$ of Type 3 calls (in O) + 1

Thus, each Type 4 call can be mapped to a unique Type 2 call, and each Type 3 call can be mapped to a unique Type 1 call (one Type 1 call remains 'unmapped'). Given the memorylessness of the call holding times, we now re-sample the residual lives of all calls using independent $\mathrm{Exp}(\mu)$ random variables such that mapped call pairs have the same residual life. This ensures that mapped calls (these belong to different systems) depart at the same time.

Type 3 Arrival: Suppose that a Type-3 arrival (into O) occurs at time s. This implies that

$$n_1(s^-) + n_2(s^-) < \tilde{n}_1(s^-) + \tilde{n}_2(s^-) = N_1 + N_2,$$

which implies that the states satisfy R2 at s^- and R1 at s. Thus, at time s, we have

- # of Type 4 calls (in O) = # of Type 2 calls (in N)
- # of Type 1 calls (in N) = # of Type 3 calls (in O)

At this point, we map each Type 2 call to a unique Type 4 call, and each Type 3 call to a unique Type 1 call. As before, we re-sample the residual service times of all calls such that mapped calls have the same residual life.

Type 4 Arrival: Suppose that a Type-4 arrival (into O) occurs at time s. This implies that

$$n_1(s^-) + n_2(s^-) < \tilde{n}_1(s^-) + \tilde{n}_2(s^-) = N_1 + N_2,$$

which implies that the states satisfy R2 at s^- and R3 at s. Thus, at time s, we have

- # of Type 4 calls (in O) = # of Type 2 calls (in N) + 1
- # of Type 1 calls (in N) = # of Type 3 calls (in O) + 1

Now, we map each Type 2 call to a unique Type 4 call, and each Type 3 call to a unique Type 1 call. Finally, we map the remaining (yet unmapped) Type 4 call to the remaining (yet unmapped) Type 1 call. As before, re-sample the residual life of all calls such that mapped calls have the same residual life.

Departures: Based on the above coupling rules for residual lives of calls across O and N, we see that four types of departures events are possible.

- Simultaneous departure of Type 2 call in N and Type 4 call in O: Clearly, the relationship between the states in O and N remains unaltered.
- Simultaneous departure of Type 1 call in N and Type 3 call in O: Clearly, the relationship between the states in O and N remains unaltered.
- Departure of an unmapped Type 1 call out of N: This can only happen if the states satisfy R2 just prior to the departure. The states then satisfy R1 post-departure.
- Simultaneous departure of a Type 1 call out of N and a Type 4 call out of O: This can only happen if the states satisfy R3 just prior to the departure. Clearly, the states will satisfy R1 post-departure.

This completes the argument that the states of the systems O and N remain related via R1, R2, or R3 at all times.

Note that any Type 3/4 departure out of the O system is always synchronized with a Type 1/2 departure out of the N system. This means that at all times, the cumulative departures out of the N system exceed the cumulative departures out of the O system. Moreover, since there is a positive rate associated with 'solo' Type 1 departures out the N system, the statement of the lemma follows.

We are now ready to complete the proof of Statement 3. Given Lemma 12, it suffices to show that for $k_1 \in \{0, 1, \cdots, N_1 - 1\}$ and $k_2 \in [0, N_2]$, $B^{(bo)}_{overall}(k, k_2)$ is strictly decreasing over $k \in [k_1, k_1 + 1]$. From our gradient calculations, it is not hard to see that $k \in [k_1, k_1 + 1]$,

$$\frac{\partial B^{(bo)}_{overall}(k, k_2)}{\partial k} = \frac{N(k_2)}{(G'(k, k_2))^2}.$$

Note that the numerator does not depend on k. It thus suffices to prove that $N(k_2) < 0$. From Lemma 12, invoking the mean value theorem, it follows that there exists $k' \in (k_1, k_1 + 1)$ such that $\frac{N(k_2)}{(G'(k', k_2))^2} < 0$, which implies that $N(k_2) < 0$. This completes the proof of Statement 3.

APPENDIX B
PROOF OF LEMMA 5

We define a sharing configuration $k = (k_1, k_2)$ to be *efficient* if there does not exist a sharing configuration k' such that $B_i(k') \leq B_i(k)$ for all i, and $B_i(k') < B_i(k)$ for some i. Let $\mathcal{P}$ denote the set of efficient configurations. Under this notation, the set $\hat{\mathcal{P}}$ of Pareto-efficient configurations is given by $\hat{\mathcal{P}} = \mathcal{P} \cap \mathcal{Q}$, where $\mathcal{Q}$ denotes the set of QoS-stable configurations.

The first step of the proof is to show that $\mathcal{P} = \mathcal{X} \setminus \mathcal{X}^o$. Note that Lemma 4 implies that there are no efficient sharing configurations in $\mathcal{X}^o$. Thus, it only remains to show that any $\hat{k} \in \mathcal{X} \setminus \mathcal{X}^o$ is efficient. For the purpose of obtaining a contradiction, suppose that $\hat{k} \in \mathcal{X} \setminus \mathcal{X}^o$ is not efficient. Then there exists $\bar{k} \in \mathcal{X}$ such that $B_i(\bar{k}) \leq B_i(\hat{k})$ for all i. From the monotonicity of the blocking probabilities along $\mathcal{X} \setminus \mathcal{X}^o$ (Statements 1 and 2 of Theorem 1), it follows that $\bar{k} \notin \mathcal{X} \setminus \mathcal{X}^o$, which implies that $\bar{k} \in \mathcal{X}^o$. Now, define for $i \in \{1, 2\}$,

$$g_i(k_1, k_2) := \max(0, B_i(\bar{k}_1, \bar{k}_2) - B_i(k_1, k_2)).$$

Consider the optimization: $\max_{k \in \mathcal{X}} g_1(k_1, k_2) g_2(k_1, k_2)$. Since this is the maximization of a continuous function over a compact domain, a maximizer $k^* \in \mathcal{X}$ exists. Moreover, the optimum value is strictly positive (follows from Lemma 4), $k^* \in \mathcal{X} \setminus \mathcal{X}^o$, and $B_i(k^*) < B_i(\bar{k})$ for all i. Thus, we have $\hat{k}, k^* \in \mathcal{X} \setminus \mathcal{X}^o$ such that $B_i(k^*) < B_i(\hat{k})$ for all i. However, this contradicts the strict monotonicity of the blocking probabilities over $\mathcal{X} \setminus \mathcal{X}^o$. Thus, we conclude that $\hat{k}$ is efficient.

Having proved that $\mathcal{P} = \mathcal{X} \setminus \mathcal{X}^o$, characterizing $\hat{\mathcal{P}}$ boils down to identifying the subset of QoS-stable sharing configurations in $\mathcal{P}$. For this, consider the three cases in the statement of the lemma separately. We give the proof for Case 1 here; the proofs for Cases 2 and 3 are on similar lines and are omitted.

Case 1 ($E(N_1 + N_2, a_1 + a_2) < E(N_i, a_i)\ \forall\ i$): We have $B_1(1, 1) < E(N_1, a_1) < B_1(1, 0)$. Thus, there a unique $\hat{k}_2 \in (0, 1)$ such that $B_1(1, \hat{k}_2) = E(N_1, a_1)$. It is easy to see that the set of sharing configurations in $\mathcal{P}$ where Provider 1 strictly improves upon its standalone blocking probability is given by $\{(y, 1) : \ y \in [0, 1]\} \cup \{(1, y) : \ y \in (\hat{k}_2, 1]\}$. Similarly, $B_2, (1, 1) < E(N_2, a_2) < B_2(0, 1)$. Thus, there is a unique $\hat{k}_1 \in (0, 1)$ satisfying $B_2(\hat{k}_1, 1) = E(N_2, a_2)$. As before, the set of sharing configurations in $\mathcal{P}$ where Provider 2 strictly improves upon its standalone blocking probability is given by $\{(y, 1) : \ y \in (\hat{k}_1, 1]\} \cup \{(1, y) : \ y \in [0, 1]\}$. Thus, the subset of QoS-stable sharing configurations in $\mathcal{P}$ is the intersection of the above sets.

REFERENCES

[1] D. Kumar, D. Manjunath, and J. Nair, "Spectrum sharing: How much to give," in *Proc. COMSNETS*, Jan. 2016, pp. 1–8.

[2] S. Sarkar, C. Singh, and A. Kumar, "A coalitional game model for spectrum pooling in wireless data access networks," in *Proc. Inf. Theory Appl. Workshop*, Jan./Feb. 2008, pp. 310–319.

[3] A. Aram, C. Singh, S. Sarkar, and A. Kumar, "Cooperative profit sharing in coalition based resource allocation in wireless networks," in *Proc. IEEE Infocom*, Apr. 2009, pp. 2123–2131.

[4] F. J. P. Karsten, M. Slikker, and G.-J. van Houtum, "Analysis of resource pooling games via a new extension of the erlang loss function," *Oper. Res.*, vol. 63, no. 2, pp. 476–488, 2015.

[5] A. Mandelbaum and M. I. Reiman, "On pooling in queueing networks," *Manage. Sci.*, vol. 44, no. 7, pp. 971–982, Jul. 1998.

[6] S. Anily and M. Haviv, "Cooperation in service systems," *Oper. Res.*, vol. 58, no. 3, pp. 660–673, May/Jun. 2010.

[7] U. Özen, M. I. Reiman, and Q. Wang, "On the core of cooperative queueing games," *Oper. Res. Lett.*, vol. 39, no. 5, pp. 385–390, Sep. 2011.

[8] *Guidelines on Spectrum Sharing*, Telecom Regulatory Authority of India, New Delhi, India, Jul. 2014.

[9] F. P. Kelly, "Loss networks," *Ann. Appl. Probab.*, vol. 1, no. 3, pp. 319–378, Aug. 1991.

[10] D. E. Everitt and N. W. Macfadyen, "Analysis of multicellular mobile radio telephone systems with loss," *Brit. Telecommun. Technol. J.*, vol. 1, no. 2, pp. 37–45, 1983.

[11] F. Kelly, "Stochastic models of computer communication systems," *J. Roy. Stat. Soc., Ser. B*, vol. 47, no. 3, pp. 379–395, Jul. 1985.

[12] R. Schassberger, "Two remarks on insensitive stochastic models," *Adv. Appl. Probab.*, vol. 18, no. 3, pp. 791–814, Sep. 1986.

[13] A. Nandigam, S. Jog, D. Manjunath, J. Nair, and B. J. Prabhu, "Sharing within limits: Partial resource pooling in loss systems," 2018, *arXiv:1808.06175*. [Online]. Available: https://arxiv.org/abs/1808.06175

[14] R. B. Myerson, *Game theory*. Cambridge, MA, USA: Harvard Univ. Press, 2013.

[15] P. J.-J. Herings and A. Predtetchinski, "Bargaining with non-convexities," *Games Econ. Behav.*, vol. 90, pp. 151–161, Mar. 2015.

[16] J. F. Nash, "The bargaining problem," *Econometrica*, vol. 18, no. 2, pp. 155–162, Apr. 1950.

[17] E. Kalai and M. Smorodinsky, "Other solutions to Nash's bargaining problem," *Econometrica*, vol. 43, no. 3, pp. 513–518, May 1975.

[18] E. Kalai, "Proportional solutions to bargaining situations: Interpersonal utility comparisons," *Econometrica*, vol. 45, no. 7, pp. 1623–1630, Oct. 1977.

[19] W. Thomson, "Nash's bargaining solution and utilitarian choice rules," *Econometrica*, vol. 49, no. 2, pp. 535–538, Mar. 1981.

[20] Adam Dutcher. (2017). *Capacity in the Cell Signal Oriented World*. [Online]. Available: https://blog.surecall.com/capacity-in-the-cell-signal-oriented-world/

[21] ITU.2. (2016). *End-to-end Quality of Service for Voice over 4G Mobile Networks, Recommendation G.1028*. [Online]. Available: https://www.itu.int/rec/T-REC-G.1028

[22] Upasana Jain. (2016). *TRAI Pushes for Powers to Penalize Telcos for Call Drops*. [Online]. Available: https://www.livemint.com/Industry/jGdVYXmyLU421aaHUBYdEP/Trai-says-most-telecom-firms-fail-to-meet-call-drops-benchma.html,

[23] A. Mukhopadhyay, A. Karthik, R. R. Mazumdar, and F. Guillemin, "Mean field and propagation of chaos in multi-class heterogeneous loss models," *Perform. Eval.*, vol. 91, pp. 117–131, Sep. 2015.

[24] S. Halfin and W. Whitt, "Heavy-traffic limits for queues with many exponential servers," *Oper. Res.*, vol. 29, no. 3, pp. 567–588, Jun. 1981.

[25] R. Atar, "A diffusion regime with nondegenerate slowdown," *Oper. Res.*, vol. 60, no. 2, pp. 490–500, Apr. 2012.

[26] A. Jensen, E. Brockmeyer, and H. L. Halstrom, "The life and works of AK Erlang," *Trans. Danish Acad. Tech. Sci.*, vol. 2, pp. 190–192, 1948.

[27] D. L. Jagerman, "Some properties of the Erlang loss function," *Bell System Tech. J.*, vol. 53, no. 3, pp. 525–551, Mar. 1974.

[28] S. Borst, A. Mandelbaum, and M. I. Reiman, "Dimensioning large call centers," *Oper. Res.*, vol. 52, no. 1, pp. 17–34, Feb. 2004.

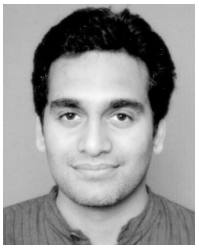
[29] F. Kelly, "Routing in circuit-switched networks: Optimization, shadow prices and decentralization," *Adv. Appl. Probab.*, vol. 20, no. 1, pp. 112–144, Mar. 1988.

[30] M. I. Reiman, "Asymptotically optimal trunk reservation for large trunk groups," in *Proc. IEEE CDC*, Dec. 1989, pp. 2536–2541.

Anvitha Nandigam received the bachelor's degree in electronics design and manufacturing from IIITDM Kancheepuram in 2014, and the master's degree in electrical engineering specializing in communications engineering from IIT Bombay in 2017. She is currently with Qualcomm Technologies, working on 5G cellular network technology. Her interests are in queuing systems and communication networks.

Suraj Jog received the B.Tech. degree in electrical engineering from IIT Bombay in 2016. He is currently pursuing the Ph.D. degree with the Electrical and Computer Engineering Department, University of Illinois Urbana–Champaign (UIUC), and his thesis is being advised by Prof. H. Hassanieh. He is interested in the area of wireless networking and RF systems, and his research is focused on investigating new networking paradigms and system designs for millimeter-wave (60 GHz) communications. He is also interested in the area of IoT and low-power wide area networks (LPWANs) and has worked on designing robust networking solutions for new application domains, such as precision agriculture and smart city monitoring. His research has been recognized with the Rambus Computer Engineering Fellowship and the M.E. Van Valkenburg Graduate Research Award at UIUC. He also won the First Position in the ACM SIGCOMM'18 Student Research Competition.

D. Manjunath received the B.E. degree from Mysore University in 1986, the M.S. degree from IIT Madras in 1989, and the Ph.D. degree from the Rensselaer Polytechnic Institute in 1993. He was with the Corporate R&D Center, General Electric, Scehenectady, NY, USA, in 1990, with the Computer and Information Sciences Department, University of Delaware, from 1992 to 1993, with the Computer Science Department, University of Toronto, from 1993 to 1994, and with the Department of Electrical Engineering, IIT Kanpur, from 1994 to 1998. He was the Head of the Computer Centre, IIT Bombay, from 2011 to 2015. He has been with the Electrical Engineering Department, IIT Bombay, since 1998, where he is currently an Institute Chair Professor. He has coauthored two textbooks *Communication Networking: An Analytical Approach* (Morgan-Kaufman Publishers, 2004) and *Wireless Networking* (Morgan-Kaufman Publishers, 2008). His research interests are in the general areas of communication networks and performance analysis. His recent research has concentrated on random networks with applications in wireless and sensor networks, network pricing, and queue control. He was a recipient of the Best Paper Award at ACM SIGMETRICS 2010. He was the TPC Chair of COMSNETS 2011 and NCC 2015 and the General Chair of ACM MobiHoc 2013 and COMSNETS 2015. He is an Associate Editor of IEEE/ACM TRANSACTIONS ON NETWORKING, *Queueing Systems: Theory and Applications*, and *Sadhana: The Proceedings of the Indian Academy of Sciences*.



Jayakrishnan Nair received the B.Tech. and M.Tech. degrees from IIT Bombay in 2007 and the Ph.D. degree from the California Institute of Technology in 2012, all in electrical engineering (EE). He has held post-doctoral positions at California Institute of Technology and Centrum Wiskunde & Informatica. He is currently an Assistant Professor in EE with IIT Bombay. His research focuses on modeling, performance evaluation, and design issues in queuing systems and communication networks.

Balakrishna J. Prabhu received the Ph.D. degree from INRIA Sophia Antipolis, France, in 2005, and the M.Sc. (Engg.) degree from IISc, India. He did post-doctoral stints at VTT, Finland, CWI, Eurandom, and TU/e, The Netherlands. He is currently a CNRS Researcher with LAAS-CNRS, Toulouse, France. His research interests include the performance analysis of communication systems using stochastic modeling and game theory.