Forty-Eighth Annual Allerton Conference
Allerton House, UIUC, Illinois, USA
September 29 - October 1, 2010

Revenue Maximization in Reservation-based Online Advertising Through Dynamic Inventory Management

Ana Radovanović
Google, Inc.
New York, NY 10011
Email: anaradovanovic@google.com

Assaf Zeevi
Columbia Business School
New York, NY 10025
Email: assaf@gsb.columbia.edu

***Abstract*—A widely used model in the online advertising industry is one where advertisers pre-purchase a reservation package of online inventory on content sites owned by publishers (e.g., CNN, amazon, etc.). Sales representatives, acting on behalf of publishers, sell inventory (impression) bundles of various types (text, video, multimedia, etc.) while trying to meet advertisers' expectations. The current process of sales is usually ad hoc and oftentimes a publisher uncontrollably runs out of a highly desirable inventory type, failing to meet the demand of his/her more valuable customers (advertisers). In this specific framework of display advertising, we propose a mathematical model for this problem and design a simple and easy to implement online impression allocation policy with provably revenue maximizing performance. Our results represent fundamental extensions to the existing theory of loss networks, given that this new application introduces novel mathematical assumptions and operational constraints.**

I. Introduction

A widely used model in online advertising industry is the one in which advertisers pre-purchase a reservation package of online inventory on content sites owned by the so called publishers (for example, CNN, amazon.com, etc.). This package consists of specified inventory bundles of various types (e.g., display, text, video, pop-up) that are priced differently. In this context, inventory is counted in units known as impressions, i.e., the number of times a certain ad appears on a Web page when users access it. Inventory categories differ in their properties, such as size, type (display, video, etc.), position, as well as monitored measures of their effectiveness (one very common example is a Click Through Rate - CTR). When online advertisers arrive to a publisher, they have a daily budget, a desirable duration of the specific advertising campaign and a performance goal, i.e., some target 'effectiveness' of a purchased package of impressions. Given the requirements and the current inventory availability, a sales person (agent) allocates a bundle of impressions for the incoming advertiser. The focus of this paper is a design of a simple, easy to implement, online inventory allocation policy for which we prove its near optimal long run performance.

The underlying dynamics of the described application bears some similarities to bandwidth sharing in communication networks. The elaborate studies in the somewhat related context of loss networks started in early '80 and have expanded to other domains encompassing pricing, congestion control and bandwidth planning (see, [6], [7], [17], and a more recent work on bandwidth sharing policies in [5], including references therein). Recently, this work found some new applications in the field of resource (workforce) management (see [13], [9] and [10]). Despite the aforementioned similarities, there are intrinsic characteristics that make the problem of impression allocations in online advertising novel from the modeling and analysis perspective. The key difference is a random budget which translates into a random inventory demand. The other, very important, property is that online advertisers do not ask for specific a resource when purchasing the bundle of impressions. What they are looking for is some notion of "effectiveness" or "quality" they will experience for the budget they invest in their advertising campaign. On the other hand, the most effective inventory is scarce and it is of huge importance for publishers to design inventory allocation schemes that will keep their clients (advertisers) satisfied, and which could potentially increase their revenue.

The policy we will propose relies on the solution of a suitably constructed linear program. This provides both guidance towards the said policy and plays a central role in establishing performance bounds. Rather surprisingly, we show that the LP solution itself is *not* sufficient for "good" planning of the online impression allocation process. In order to amortize the uncertainty of the incoming and overflowing demand, one must incorporate suitable "safety stocks" of impressions designated to handle demand overflows. We introduce a fundamentally new techniques that allow us to properly size these safety stocks and hence allowing us to prove the policy performance is asymptotically optimal.

II. Problem Formulation

Let a stream of M classes of online advertisers arrive to a publisher according to independent Poisson processes in time. Assume that advertisers of class $1 \leq m \leq M$ arrive at time points $\{\tau_n^{(m)}\}$, with rate $\lambda_m = (\mathbb{E}[\tau_n^{(m)} - \tau_{n-1}^{(m)}])^{-1} > 0$. An advertiser of type m, $1 \leq m \leq M$, arriving at time $\tau_n^{(m)}$, brings some daily budget he is willing to spend on impressions

(online inventory) that we model by a random variable $B_n^{(m)}$, $B_n^{(m)} < B < \infty$; we assume that $\{B_n^{(m)}\}$ are mutually independent and independent from $\cup_m \{\tau_j^{(m)}\}$.

We assume that there are $K < \infty$ inventory (impression) types, with the corresponding prices per impression p_k, $1 \leq k \leq K$. Price is significantly related to 'effectiveness' of the corresponding inventory. (For *direct response* advertisers who care about clicks, effectiveness is usually measured through the Click-Through-Rate (CTR), which represents the proportion of impressions that are bought and result in a click.) If an advertiser arriving at time $\tau_n^{(m)}$ ends up buying a package $(I_{n,1}^{(m)}, \ldots, I_{n,K}^{(m)})$ with $I_{n,1}^{(m)}$ impressions of type 1, ..., and $I_{n,K}^{(m)}$ impressions of type K, the resulting effectiveness, say $\Delta(I_{n,1}^{(m)}, \ldots, I_{n,K}^{(M)})$, is

$$\Delta(I_{n,1}^{(m)}, \ldots, I_{n,K}^{(m)}) = \varepsilon_1 I_{n,1}^{(m)} + \cdots + \varepsilon_K I_{n,K}^{(m)}, \tag{1}$$

where we use $\varepsilon_k > 0$, $1 \leq k \leq K$, to denote the effectiveness of inventory k. (Here, we assume that effectiveness is some given and fixed parameter.) When sold to an advertiser of class m, impressions are reserved for some random amount of time with cumulative distribution function $G^{(m)}$ and expected value $\mu^{(m)} \triangleq \int_0^\infty (1 - G^{(m)}(u))du$. On the other hand, daily impression capacities are finite, i.e., the publisher can provide only a finite amount of impressions of type $1 \leq k \leq K$, say $C_k = \phi_k C$, $0 < \phi_k \leq 1$, per day. The process of inventory allocation on behalf of the publisher is conducted by an assigned sales person. When an advertiser arrives with a daily budget $B_n^{(m)}$, a sales person allocates inventory so that $B_n^{(m)} = p_1 I_{n,1}^{(m)} + \cdots + p_K I_{n,K}^{(m)}$. Each sales agent's goal is to create a package of impressions that will utilize the invested budget in a "preferred" way for the class of advertisers it is sold to. We assume that these target budget utilizations, i.e., proportions of the budget that turn out to be effective (e.g., turn into clicks), are given by ε_m^*, $1 \leq m \leq M$, for the corresponding advertiser class. Given that a sales agent sells a package of impressions $(I_1^{(m)}, \ldots, I_K^{(m)})$ to an advertiser of class m and budget $B_n^{(m)}$, the proportion of the budget that gets utilized is

$$E_n^{(m)} = \frac{\varepsilon_1 p_1 I_{n,1}^{(m)} + \cdots + \varepsilon_K p_K I_{n,K}^{(m)}}{B_n^{(m)}}.$$

An important constraint in the allocation process is having "small" deviations from the target package effectiveness.

Our objective in this paper is to design a dynamic inventory allocation policy that maximizes expected long run daily revenue subject to constraints on deviations from the target allocation effectiveness, i.e. $|E_n^{(m)} - \varepsilon_m^*|$, and inventory capacity constraints. This is a constrained stochastic control problem, which even in the simplified Markovian framework (exponentially distributed campaign durations) with fixed budget values, and using dynamic programming principles, becomes intractable due to the high dimensionality of the state space. Instead of following that path, we first relax the original problem formulation and concentrate on maximizing the *expected net revenue rate* that we define as

$$\max \mathbb{E}\left\{ \sum_{m=1}^{M} \sum_{i \in \mathcal{N}_n^{(m)}} B_i^{(m)} - \sum_{m=1}^{M} \sum_{i \in \mathcal{N}_n^{(m)}} B_i^{(m)} \gamma^{(m)} (\varepsilon_m^* - E_i^{(m)}) \right\}, \tag{2}$$

where parameters $\gamma^{(m)}$, $1 \leq m \leq M$, allow the publisher to differently penalize 'under-performance' among distinct advertiser classes based on the previous experience, revenue goals, market focus plans, etc. We use $\mathcal{N}_n^{(m)}$ in (2) to denote a set of still-active transactions of type m at the moment τ_n. Next, we concentrate on the relaxed objective which, in conjunction with the inventory capacity constraints, we solve in an approximate manner. We use a knapsack-type linear program (LP) (explained in Section III) that provides an upper bound on expected net revenue rate of any stationary, state-dependent, inventory allocation policy. Furthermore, we use the solution of the constructed LP to design a simple online allocation policy, which we prove is asymptotic optimality in a regime where advertiser arrival rates (demand) and inventory capacities grow large (discussed in Sections III and IV and VI).

III. Stationary Allocation Policies

We focus on a set of stationary policies that upon an arrival of an advertiser, make impression allocation decisions based on her class, budget, utilization targets and the current impression availability. For each policy π, let $\mathscr{R}_\pi^{\bar{C}}(T)$ be the net revenue rate achieved by policy π over interval $[0,T]$ with inventory capacities $\bar{C} \triangleq (C_1, \ldots, C_K)$. Then, we define the expected long-run net revenue rate of a policy π as

$$\mathscr{R}^{\bar{C}}(\pi) \triangleq \lim_{T \to \infty} \frac{\mathbb{E}_\pi[\mathscr{R}_\pi^{\bar{C}}(T)]}{T}, \tag{3}$$

where the expectation $\mathbb{E}_\pi$ is taken with respect to probability measure induced by policy π.

At any time point t, the state of the system is specified by the class of an advertiser with its campaign in process, the amount of impressions of each inventory engaged by the specific campaign, as well as the time that elapsed from the moment of the corresponding advertiser/campaign arrival. Since we focus on state-dependent policies, they can be represented as measurable functions from the state-space defined above to actions. Note that each state-dependent policy induces a Markov process over the state-space. Then, using analogous arguments as in the Appendix of [11] (which is the extended version of [12]), we prove the specific version of Theorem 1 from [15], and show the existence of a unique stationary distribution which is ergodic. Thus, for each state-dependent policy π, the limit in (3) is well-defined and the expectation in the numerator of (3) exists.

In this paper, we concentrate on a specific subset of state-dependent policies that allocate class-dependent proportions of advertiser's budget on available inventory. In the case where there is not enough of policy-prescribed inventory available at the moment of an advertiser arrival, we propose a way to substitute this inventory with other inventory types.

A. Optimal Online Allocation Proportions

In this section we use a simple linear program (LP) that will allow us to: (1) estimate an upper bound on the achievable long-run revenue rate, and (2) provide the policy design guidelines. The LP exploits the ergodicity discussed in the previous section and uses the existence of the long-run averages to compute the optimal class-dependent *static* impression allocation proportions. We use these proportions to design an online policy which we prove achieves the optimal long-run net revenue rate in the regime where advertiser arrival rates and capacities grow large.

What makes the dynamics in this paper intrinsically different from that of many similar loss network models is that all online inventory is substitutable and advertiser requirements are expressed not in terms of the amount of inventory, but its target performance (i.e., desired effectiveness). Note that revenue penalties (as described at the end of Section II) in missing advertiser target effectiveness depend significantly on the choice of inventory substitutes, which provides the key intuition for the design of the well-performing online allocation policy.

As we discussed before, we concentrate on a set of state-dependent policies where any such policy induces a Markov process on the state-space of the system with a unique stationary distribution that is ergodic. In particular, for each advertiser class m, inventory k and a given state-dependent policy π, there exists a stationary proportion $\alpha_{m,k}^{(\pi)}$ of its budget that is spent on inventory k, which is at the same time equal to the long-run proportion of advertiser m budgets spent on inventory k. Thus, any state-dependent policy π is associated with the stationary proportions $\alpha_{m,k}^{(\pi)}$, $1 \le m \le M$, $1 \le k \le K$. Furthermore, using Little's law, we can use the stationary proportions $\alpha_{m,k}^{(\pi)}$ to express the average amount of resource k engaged by advertisers of type m as $\lambda^{(m)}\mu^{(m)}\mathbb{E}B^{(m)}\alpha_{m,k}^{(\pi)}/p_k = \rho^{(m)}\mathbb{E}B^{(m)}\alpha_{m,k}^{(\pi)}/p_k$. (Note that $\rho^{(m)} = \lambda^{(m)}\mu^{(m)}$ is the expected number of class m advertisers being served in the system assuming that there is always some inventory, potentially very ineffective, that a sales person can use to create a package to sell.) Thus, in conjunction with (2), it follows that the expected long-run net revenue rate of a policy π can be expressed as

$$\sum_{m=1}^{M} \rho^{(m)}\mathbb{E}B^{(m)} - \sum_{m=1}^{M} \rho^{(m)}\mathbb{E}B^{(m)}\gamma^{(m)}(\varepsilon_m^* - \bar{\alpha}_m^{(\pi)} \times \bar{\varepsilon}),$$

where $\bar{\alpha}_m^{(\pi)} \triangleq (\alpha_{m,1}^{(\pi)}, \ldots, \alpha_{m.K}^{(\pi)})$, $\bar{\varepsilon} \triangleq (\varepsilon_1, ..., \varepsilon_K)$, and $\times$ denotes the scalar product of the two vectors. The second term follows from $\mathbb{E}E_n^{(\pi,m)} = \sum_{k=1}^{K} \alpha_{m,k}^{(\pi)}\varepsilon_k = \bar{\alpha}_m^{(\pi)} \times \bar{\varepsilon}$.

The physical constraints of the system that we analyze in this paper imply that, for any feasible allocation policy, it is not possible to find more than C_k, $1 \le k \le K$, impressions being allocated by ongoing ad campaigns. Therefore, in view of the notation from the above, the expected amount of impressions k allocated by active campaigns satisfies

$$\sum_{m=1}^{M} \rho^{(m)}\mathbb{E}B^{(m)}\alpha_{m,k}^{(\pi)}/p_k \le C_k.$$

An intrinsic property of the inventory allocation dynamics in this paper relates to the ways of 'prioritizing' the choice of inventory assigned to a specific class of advertisers. More specifically, our goal is to design a budget-proportion rule that would be a guideline for sales people on how to allocate the budget of an incoming advertiser. Furthermore, in the case when there is not enough available suggested inventory (i.e., already engaged by the ongoing ad campaigns), there needs to be a guideline on how to choose the "right" impression substitutions. Thus, inventory $1 \le k \le K$ utilization is affected by the two sources of demand: (i) the demand that is instantaneously allocated using the policy prescribed budget proportion, and (ii) forwarded demand generated in the case where there is not enough of available policy-recommended inventory and the specific inventory is used as a substitution. We design the rule that uses the prescribed budget proportions to allocate the incoming advertiser requests. In the case this demand can not be met due to limited resource availability, the rule uses a specific substitution, as described in Subsection III-B.

Next, we introduce the linear program (LP) and use its solution to construct a simple online impression allocation rule that we call *Waterfall Allocation* (WA) policy. In view of the previous discussion in this section, we focus on solving the following LP relaxation that maximizes the long-run expected net revenue rate:

$$\max \sum_{m=1}^{M} \rho^{(m)}\mathbb{E}B^{(m)} - \sum_{m=1}^{M} \rho^{(m)}\mathbb{E}B^{(m)}\gamma^{(m)}(\varepsilon_m^* - \bar{\alpha}_m^{(\pi)} \times \bar{\varepsilon}) \quad (4)$$

$$\text{s.t.} \sum_{m=1}^{M} \rho^{(m)}\mathbb{E}B^{(m)}\alpha_{m,k}^{(\pi)}/p_k \le \hat{C}_k, \ 1 \le k \le K, \quad (5)$$

$$0 \le \alpha_k^{(m)} \le 1, \ 1 \le k \le K, \ 1 \le m \le M; \quad (6)$$

Note that we use $\hat{C}_k \le C_k$ in the capacity constraint (5) instead of the total capacity C_k for inventory k. The role of this, *adjusted capacity value*, is to amortize the variability in the 'demand' by using $\Delta C_k \triangleq C_k - \hat{C}_k$ to meet the demand of the forwarded traffic (in the case the original inventory choice can not be met due to the non-availability). More explicit details about the policy that we propose are presented in the next subsection.

The LP in (4) solves for the optimal values of $\bar{\alpha}_m^{(\pi)}$, $1 \le m \le M$, and is equivalent to

$$\max \sum_{m=1}^{M} \rho^{(m)}\mathbb{E}B^{(m)}\gamma^{(m)}\bar{\alpha}^{(m)} \times \bar{\varepsilon} \quad (7)$$

$$\text{s.t.} \sum_{m=1}^{M} \rho^{(m)}\mathbb{E}B^{(m)}\alpha_k^{(m)} \le \hat{C}_k p_k, \ 1 \le k \le K. \quad (8)$$

$$0 \le \alpha_k^{(m)} \le 1, \ 1 \le k \le K, \ 1 \le m \le M. \quad (9)$$

Then, the solution of the previous linear program is obtained from the following lemma.

Lemma 1: Assume that advertiser classes are enumerated in the decreasing order of $\gamma^{(m)}$, i.e.,

$$\gamma^{(1)} \ge \gamma^{(2)} \ge \ldots, \quad (10)$$

and that inventories are enumerated in the decreasing order of ε_k, i.e.,

$$\varepsilon_1 \geq \varepsilon_2 \geq \dots. \tag{11}$$

Then, the solution of the LP (7), $\bar{\alpha}^*_m$, $1 \leq m \leq M$, has a simple structure described as follows:

1) Start from $m = 1$ and $k = 1$.
2) Keep setting $\alpha^*_{m,1} = 1$, $m = 1, 2, \dots, m^*(1)$, (corresponds to spending all of the expected budget of class m on inventory 1) until the violation of the capacity constraint, i.e., $\sum_{m=1}^{m^*(1)} \rho^{(m)} \mathbb{E}B^{(m)} > \hat{C}_1 p_1$. Note that, in general, class $m^*(1)$ has the solution $\alpha^*_{m^*(1),1} \in [0,1]$, obtained by solving $\sum_{m=1}^{m^*(1)-1} \rho^{(m)} \mathbb{E}B^{(m)} + \rho^{(m^*(1))} \mathbb{E}B^{(m^*(1))} \alpha^*_{m^*(1),1} = \hat{C}_1 p_1$.
3) Then, start from class $m = m^*$, and allocate $(1 - \alpha^*_{m^*(1),1}) \mathbb{E}B^{(m^*(1))}$, $\mathbb{E}B^{(m^*(1)+1)}$, ... budgets similarly as before till capacity $\hat{C}_2$ is exhausted, etc.
4) Continue the same process until we exhaust the expected budget of all M classes of advertisers.

Proof: It is straightforward to show the previous claim by starting from any feasible point $\bar{\alpha}^{(m)}$, $1 \leq m \leq M$, and proving that reallocating budgets towards assignments where more effective inventory is delivered to more penalizing advertisers leads to the larger objective value in (7). In this regard, in view of the orderings in (10) and (11), pick any $\bar{\alpha}^{(m)} \neq \bar{\alpha}^*_m$, $1 \leq m \leq M$. Then, there exist advertiser classes $m < n$ and inventories $i < j$ such that

$$\alpha^{(m)}_j > 0 \text{ and } \alpha^{(n)}_i > 0. \tag{12}$$

Without loss of generality, we assume that feasible proportions $\bar{\alpha}^{(m)}$, $1 \leq m \leq M$, incorporate the scenario where there is available inventory n, in which case, we assign it to some 'imaginary' advertiser with $\gamma^{(M+1)} = 0$. Next, if $b \triangleq \min(\rho^{(m)} \mathbb{E}B^{(m)} \alpha^{(m)}_j, \rho^{(n)} \mathbb{E}B^{(n)} \alpha^{(n)}_i)$, then, by exchanging inventories i and j worth budget b between advertisers m and n, the objective value increases by

$$b\gamma^{(m)}(\varepsilon_i - \varepsilon_j) + b\gamma^{(n)}(\varepsilon_j - \varepsilon_i) = b(\gamma^{(m)} - \gamma^{(n)})(\varepsilon_i - \varepsilon_j) > 0.$$

By continuing this reallocation procedure in a similar manner, we reach the optimal solution that follows from (1-4) in the statement of this lemma. ♢

Remark 1: By analyzing the corresponding Lagrange dual formulation, it is straightforward to check that the coefficients for which the objective function reaches the maximum value are those that maximize products $\gamma^{(m)} \varepsilon_k$, which implies the solution we state in Lemma 1.

Remark 2: Note that the LP in (7) enforces the capacity constraint (8) only in expectation, while in the original problem, this constraint has to hold for every sample path. It follows that the LP defined by (7) - (9) represents a relaxation and provides an upper bound on the optimal expected long-run revenue rate that can be 'earned' from capacities $\hat{C}_1, \dots, \hat{C}_K$.

B. Waterfall Allocation Policy

Finally, we propose the *Waterfall Allocation* (WA) policy, that intuitively follows from the solution of the LP above. Define $\mathcal{K}(m)$ to be a set of indexes of inventories $1 \leq k \leq K$ for which $\alpha^{(m)}_k > 0$. Using the LP solution from above, it is natural to suggest the following dynamic impression allocation policy:

- Assume that there is an arrival of an advertiser of type m at time $\tau^{(m)}_n$ with budget $B^{(m)}_n$.
- Then, with probability $\alpha^{(m)}_k$, a sales agent chooses inventory k to allocate $B^{(m)}_n / p_k$ impressions out of $\hat{C}_k$ impressions used for meeting the 'external' demand. If there are enough available impressions k, the sales agent is done.
- Otherwise, in the case there are no enough available impressions of type k, i.e., $\sum_{m=1}^{M} \sum_{i \in \mathcal{N}^{(m)}_{n,k}} B^{(m)}_i + B^{(m)}_n > \hat{C}_k p_k$, a sales agent keeps assigning $l = \max_k\{k : k \notin \mathcal{K}(m)\} + 1$, $l+1$, ..., until all of the budget $B^{(m)}_n$ is exhausted. In meeting this, 'forwarded', demand, the sales agent first tries the reserved 'safety stock' of size $C_l - \hat{C}_l$, then, in the case there are no enough available resources, the agent tries allocating from $\hat{C}_l$ impressions, after which, if necessary, she/he moves to inventory $l+1$, etc.

More descriptively, the WA policy tries to assign the LP prescribed inventory with probabilities corresponding to $\bar{\alpha}^{(m)}$, $1 \leq m \leq M$, if available. If there are not enough of this inventory, it allocates less effective inventories. Therefore, the demand that can not be met by a prescribed inventory pool overflows to 'other', less effective, inventory pools until the whole budget is exhausted. The whole process resembles a waterfall. This overflow effect is the reason for having tighter capacity constraints in (8). Having reserved 'safety stocks' with the exclusive purpose of meeting the forwarded demand amortizes overflow effects and reduces the deviation from the mean-based capacity planning obtained from (7). More specifically, due to the stochastic nature of the process of advertisers' demand, inventory availability could significantly fluctuate and deviate from the optimal static value, especially since the demand that can not be met by the LP prescribed inventory is forwarded to other inventory pools. This incures potentially large loss of the net revenue rate.

IV. Asymptotic Optimality

In this section we state and discuss the main result of this paper. Our goal is to show that under careful sizing of the *safety stocks* $\Delta C_k \triangleq C_k - \hat{C}_k$, $1 \leq k \leq K$, the performance of the WA policy approaches the optimal in the regime where inventory capacities $C_k = \phi_k C$, $1 \leq k \leq K$, and arrival rates, λ_m, $1 \leq m \leq M$, grow large with the same rate r.

The WA policy uses allocation proportions obtained from LP (7) in making allocation decisions. The assignment recommendations are done based on deterministic, mean value properties of incoming demand and using effective capacities $\hat{C}_k$, $1 \leq k \leq K$, for each inventory. In general, given the intrinsic

variability of the incoming demand, the actual inventory usage can exceed the effective capacity, without violating the capacity constraint (8). Thus, as explained before, (7) provides an upper bound for the expected long-run net revenue rate associated with any state-dependent policy π selling inventory with capacities $C_1-\Delta C_1,\ldots,C_K-\Delta C_K$. In addition, given that some extra revenue is earned from selling 'safety stocks', the expected long-run net revenue rate of state-dependent policy π, $\mathscr{R}^{\bar{C}}(\pi)$, can be upper bounded as

$$\mathscr{R}^{\bar{C}}(\pi) \le \max_{\pi} \mathscr{R}^{\hat{C}}(\pi) + (\max_k p_k) \sum_{k=1}^{K} \Delta C_k \triangleq \bar{\mathscr{R}}^{\bar{C}}(\pi), \quad (13)$$

where we use $\bar{\mathscr{R}}^{(C-\Delta C)}(\pi)$ to denote expected long-run net revenue rate of a state-dependent policy with inventory capacities $\hat{C}_1,\ldots,\hat{C}_K$.

In view of the previous definitions, let $\mathscr{R}^{\bar{C}}(\mathrm{WA})$ be the expected long-run net revenue rate of the *Waterfall Allocation* (WA) policy proposed in the previous section, and let $\mathscr{R}^{\bar{C}}_*$ be the optimal achievable long-run net revenue rate among all state dependent policies. Then, using (13), we have

$$\mathscr{R}^{\bar{C}}(\mathrm{WA}) \le \mathscr{R}^{\bar{C}}_* \le \bar{\mathscr{R}}^{\bar{C}}(\pi).$$

Proving that $\mathscr{R}^{\bar{C}}(\mathrm{WA})$ converges to the upper bound $\bar{\mathscr{R}}^{\bar{C}}(\pi)$ as $C\to\infty$, implies asymptotic optimality of the WA policy. The following theorem contains the main result of the paper.

Theorem 1: For 'safety staffing' satisfying $\Delta C_k = \kappa\sqrt{C\log C}$ with appropriatelly estimated constant $\kappa<\infty$, the WA policy achieves the optimal long-run net revenue rate when arrival rates λ_m and impression capacities C_k grow large, i.e.,

$$\frac{\mathscr{R}^{\bar{C}}(\mathrm{WA})}{\mathscr{R}^{\bar{C}}_*} \to 1 \ \text{ as } \ r\to\infty;$$

the asympotic optimality holds in the regime where transaction arrival rates $\lambda^{(m),r}$, $1\le m\le M$, and capacities C^r_k, $1\le k\le K$, grow large with a common scaling factor r ($C^r_k = rC_k$, $\lambda^{(m),r} = r\lambda^{(m)}$).

In the following sections we use $f(x)\sim g(x)$, as $x\to x_0$, to denote $\lim_{x\to x_0} f(x)/g(x) = 1$.

V. Preliminary Results

In this subsection we state two technical results that we use in Section VI to justify Theorem 1. Observe a Poisson process of transaction arrivals for a unit amount of resource sharing a common resource pool of infinite capacity. Let the Poisson rate be λ and service requirements be mutually independent and independent from the process of arrivals, as well as generally distributed with finite mean $\mu = \mathbb{E}S < \infty$. Assuming that this M/G/∞ system is in the stationary regime, the number of active transactions, say X, is Poisson with rate $\rho = \lambda\mathbb{E}S$. Then, the following asymptotic result holds:

Lemma 2: Let X be a Poisson random variable with mean ρ. Then,

$$\mathbb{P}[X > \rho + f(\rho)] \sim \bar{\Phi}_0\left(\frac{f(\rho)}{\sqrt{\rho+f(\rho)}}\right) \ \text{ as } \ \rho\to\infty,$$

where $\bar{\Phi}_0$ is the tail of the standard normal Gaussian random variable, and $f(\rho)>0$ is an increasing function of ρ, with $f(\rho)\to\infty$ as $\rho\to\infty$.

The key ingredient in the proof of our main result in Section VI extends the results by Leadbetter [8], Berman [2] and Slepian [16] to the treatment of the specific stationary Markov sequence. To this end, we observe a stationary sequence $\{X_i\}$, where X_i represents the number of active transactions in a M/G/∞ system at the moment τ_i of ith arrival. We show that given any positive finite number $L<\infty$, under mild conditions on service time distribution and a careful safety stock sizing of $f_n(\rho,L)$, the number of times process $\{X_i\}$ exceeds level $\rho+f_n(\rho,L)$ in an interval of time (τ_i,τ_{i+n}), say $E^{(i)}_n(\rho+f_n(\rho,L))$, can be upper bounded by a Poisson random variable for n large. Given that $\{X_i\}$ is stationary, without loss of generality, we can use $E_n(\rho+f_n(\rho,L)) \equiv E^{(i)}_n(\rho+f_n(\rho,L))$.

In the following analysis, we impose a mild assumption on the residual lifetime of an ongoing contract (job) in the system, say S_e, with its excess cummulative distribution function

$$G_e(t) = \mathbb{P}[S_e<t] = \frac{1}{\mu}\int_0^t G^c(u)du, \ \ t\ge 0,$$

where $G^c(t) = 1-G(t)$.

Assumption: Let the residual lifetime S_e satisfy

$$\int_0^\infty \mathbb{P}[S_e>u]^2 du < \infty.$$

Theorem 2: Under the Assumption above and for any positive finite number $L<\infty$, there exists a sequence of functions $f_n(\rho,L)$ such that the number of times the process $\{X_i\}$ exceeds $\rho+f_n(\rho,L)$ in an interval of time $[\tau_i,\tau_{i+n})$, say $E_n(\rho+f_n(\rho))$, is bounded from above by a Poisson random variable with finite rate $L<\infty$ for all n large enough.

Sketch of the proof: Due to space limitations, we outline the key elements of the proof. The result of this theorem relies on findings from [8] extended to the specific case of the sequence $\{X_i\}$ analyzed in this paper. Define $u_n(\rho,L)$ as

$$\mathbb{P}[X_1 > \rho + u_n(\rho,L)] = \frac{L}{n}. \quad (14)$$

Then, we prove the following two conditions:

1)
$$n\sum_{j=2}^{n} |\mathbb{P}[X_1>\rho+u_n(\rho,L), X_j>\rho+u_n(\rho,L)] - \mathbb{P}[X_1>\rho+u_n(\rho,L)]\mathbb{P}[X_j>\rho+u_n(\rho,L)]| \to 0 \ \text{ as } \ n\to\infty;$$

2) Any subsequence $X_{i_1},X_{i_2},\ldots,X_{i_l},X_{j_1},X_{j_2},\ldots,X_{j_s}$, $i_1<i_2<\cdots<i_l<j_1<j_2<\cdots<j_s$, obtained from $\{X_i\}$ with $|j_1-i_l|\ge k$, satisfies
$$|\mathbb{P}[X_{i_1}\le\rho+u_n(\rho,L),\ldots,X_{j_1}\le\rho+u_n(\rho,L),\ldots] - \mathbb{P}[X_{i_1}\le\rho+u_n(\rho,L),\ldots]\mathbb{P}[X_{j_1}\le\rho+u_n(\rho,L),\ldots]| \le \alpha_{n,k}\to 0,$$
as $\rho\to\infty$, $k\to\infty$;

If one can prove that the two previously stated conditions hold, then the two main results, Theorems 5.1 and 5.2, of [8]

hold, implying that $E_n(\rho+u_n(\rho,L))$ converges to the Poisson random variable with rate L as $n\to\infty$.

The proof of condition (1) relies on the fact that the conditional distribution

$$(X_j|X_1=u)$$

is asymptotically normal when u and $\rho=\mathbb{E}X_j$ grow large, with mean

$$\mu(u,\rho)\triangleq\rho+(u-\rho)G_e^c(\tau_j-\tau_1), \tag{15}$$

and variance

$$v(u,\rho)\triangleq uG_e^c(\tau_j-\tau_1)G_e(\tau_j-\tau_1)+\rho G_e(\tau_j-\tau_1); \tag{16}$$

The preceding statements are the result of Theorem 1 in [4]. Then, in view of (15) and (16), we have, as $\rho\to\infty$,

$$\begin{aligned}
&\mathbb{P}[X_j>\rho+u_n(\rho,L)|X_1>\rho+u_n(\rho,L)]\\
&\sim\sum_{u>\rho+u_n(\rho)}\mathbb{P}[X_1=u]\\
&\qquad\times\left\{\int_{\rho+u_n(\rho,L)}^{\infty}\frac{1}{\sqrt{2\pi v(u,\rho)}}e^{-(s-\mu(u,\rho))^2/2v(u,\rho)}ds\right\}\\
&\equiv\sum_{u>\rho+u_n(\rho,L)}\mathbb{P}[X_1=u]\bar{\Phi}_{\mu(u,\rho),v(u,\rho)}(\rho+u_n(\rho,L)),
\end{aligned}\tag{17}$$

where $\bar{\Phi}_{\mu(u,\rho),v(u,\rho)}(\rho+u_n(\rho,L))$ represents the tail of the normally distributed random variable with mean $\mu(u,\rho)$ and variance $v(u,\rho)$, taken at value $\rho+u_n(\rho,L)$. Then, using $v(u,\rho)\geq\rho(1-r_{j-1})^2$, $r_{j-1}\triangleq G_e^c(\tau_j-\tau_1)$, (17) and Lemma 2, we obtain, as $\rho\to\infty$,

$$\begin{aligned}
&|\mathbb{P}[X_1>\rho+u_n(\rho,L),X_j>\rho+u_n(\rho,L)]\\
&\qquad-\mathbb{P}[X_1>\rho+u_n(\rho,L)]\mathbb{P}[X_j>\rho+u_n(\rho,L)]|\\
&\lesssim|\mathbb{P}[X_1^*>\rho+\sqrt{\rho}\nu_n(\rho,L),X_j^*>\rho+\sqrt{\rho}\nu_n(\rho,L)]\\
&\qquad-\mathbb{P}[X_1^*>\rho+\sqrt{\rho}\nu_n(\rho,L)]\mathbb{P}[X_j^*>\rho+\sqrt{\rho}\nu_n(\rho,L)]|,
\end{aligned}\tag{18}$$

where X_1^* and X_j^* are normally distributed random variables with mean ρ, variance ρ, covariance $\frac{\mathbb{E}[X_1^*X_j^*]-\mathbb{E}X_1^*\mathbb{E}X_j^*}{\rho}=r_{j-1}$, and

$$\nu_n(\rho)\triangleq\frac{u_n}{\sqrt{\rho+u_n(\rho)}}. \tag{19}$$

For more reading on jointly normal random variables, an interested reader is referred to Section 6.4 of [14].

After normalizing normal random variables X_1^* and X_j^*, we derive, as $\rho\to\infty$,

$$\begin{aligned}
&|\mathbb{P}[X_1>\rho+u_n(\rho,L),X_j>\rho+u_n(\rho,L)]\\
&\qquad-\mathbb{P}[X_1>\rho+u_n(\rho,L)]\mathbb{P}[X_j>\rho+u_n(\rho,L)]|\\
&\lesssim|\mathbb{P}[\mathscr{X}_1^*>\nu_n(\rho,L),\mathscr{X}_j^*>\nu_n(\rho,L)]\\
&\qquad-\mathbb{P}[\mathscr{X}_1^*>\nu_n(\rho,L)]\mathbb{P}[\mathscr{X}_j^*>\nu_n(\rho,L)]|.
\end{aligned}\tag{20}$$

Then, using Lemma 4.3 of [8], the right hand side of (20) is bounded by

$$\begin{aligned}
&|\mathbb{P}[\mathscr{X}_1^*>\nu_n(\rho,L),\mathscr{X}_j^*>\nu_n(\rho,L)]\\
&\qquad-\mathbb{P}[\mathscr{X}_1^*>\nu_n(\rho,L)]\mathbb{P}[\mathscr{X}_j^*>\nu_n(\rho,L)]|\\
&\leq H|r_{j-1}|e^{-\frac{\nu_n(\rho,L)^2}{1+|r_{j-1}|}},
\end{aligned}$$

for all ρ large and some constant $H<\infty$. Finally, in conjunction with the Assumption, we obtain

$$\sum_{j=1}^{\infty}r_j^2\leq\int_0^{\infty}\mathbb{P}[S_e>u]^2du<\infty,$$

which is a sufficient condition for Lemma 4.3 of [8] to hold for the sequence $\{\mathscr{X}_i^*\}$ and, in conjunction with (20), for the sequence $\{X_i\}$ as well for ρ large. This completes the proof of condition (1). The proof of condition (2) is quite technical and uses different types of arguments than the rest of the paper. Due to space limitations, we omit it in this paper.

After showing that conditions (1) and (2) hold, we obtain that, for all ρ large,

$$\lim_{n\to\infty}E_n(\rho+u_n(\rho,L))\to\mathscr{P}(L), \tag{21}$$

where we use $\mathscr{P}(L)$ to denote a Poisson random variable with mean L.

Next, we discuss the relation between $\nu_n(\rho,L)$, ρ and n and relate it to $u_n(\rho,L)$. Using (14) and Lemma 2, as well as $\bar{\Phi}_0(x)\sim\frac{\phi(x)}{x}$ as $x\to\infty$ for the standard normal distribution function $\bar{\Phi}_0(x)$ and density $\phi(x)$ (see, for example, 26.2.13 of [1]), we obtain, as $n\to\infty$,

$$\frac{1}{2}\nu_n(\rho,L)^2+\log\nu_n(\rho,L)\sim-\log L+\log n-\log\sqrt{2\pi}. \tag{22}$$

Note that the asymptotic behavior of $\nu_n(\rho,L)$ does *not* depend on ρ and, without loss of generality, we could as well write $\nu_n(L)\equiv\nu(\rho,L)$. Thus, using (22), for any $\varepsilon>0$, finite $L<\infty$ and ρ, n large enough

$$(1-\varepsilon)\sqrt{2\log n}\leq\nu_n(L)\leq(1+\varepsilon)\sqrt{2\log n}. \tag{23}$$

Next, by replacing (19) in (23), and solving the inequalities with respect to $u_n(\rho,L)$, we obtain that, for any finite L, and all n and ρ large enough, and by choosing $f_n(\rho,L)$ which satisfies

$$\begin{aligned}
&f_n(\rho,L)\\
&\geq(1+\varepsilon)\log n+\sqrt{(1+\varepsilon)^2(\log n)^2+2(1+\varepsilon)\rho\log n}\\
&\triangleq\kappa(\varepsilon,\rho,n),
\end{aligned}\tag{24}$$

$$\tag{25}$$

the number of times the process $\{X_i\}$ exceeds level $\rho+f_n(\rho,L)$ is upper bounded by $E_n(\rho,u_n(\rho,L))$ for large ρ and n, which, in conjunction with (21), completes the proof of the theorem.

$\Diamond$

VI. Outline of the Proof of the Main Result

In this section, we present the outline of the analysis of the long-run performance of the WA policy proposed in Subsection III-B. In order to show that the WA is asymptotically optimal in the regime where advertiser arrival rates and inventory capacities grow with the same rate, it is enough to prove that the long-run proportion of demand that gets forwarded due to the lack of the original inventory choice converges to zero in the analyzed regime. This directly follows from the analysis of the revenue rate generated by each inventory $1 \leq k \leq K$, which equals to

$$\sum_{m=1}^{M} \alpha_k^{(m)} \mathbb{E} B_i^{(m)} / p_k (1 - p_{m,k}^{(b)}(C_k, \Delta C_k)), \tag{26}$$

where $p_{m,k}(C_k, \Delta C_k)$ represents the long run proportion of transactions m that get forwarded from inventory k. Then, if we show that for safety sizing equal to $\Delta C_k = \kappa \sqrt{C \log C}$,

$$\max_{1 \leq m \leq M} p_{m,k}(C_k, \Delta C_k) \to 0 \text{ as } C \to \infty,$$

in conjunction with (26) and (13), we obtain that the performance of the WA policy converges to its upper bound in the particular regime, which means that it is asymptotically optimal.

Let $p_k^{(b)}(C_k, \Delta C_k)$ be the long run proportion of incoming transactions to inventory k that get forwarded to other inventories. In view of the ergodic property discussed in Section III, $p_{m,k}^{(b)}(C_k, \Delta C_k)$, $1 \leq m \leq M$, is well defined and equal to the probability that an incoming transaction gets forwarded due to inability to meet its impression demand. We define $\{tau_n\} \triangleq \cup_{m=1}^{M} \{\tau_n^{(m)}\}$ and use $\mathscr{N}_{n,k}^{(m)}$, $\alpha_k^{(m)} > 0$, to denote indices of active transactions m, arrived before τ_n, that are taking impressions k at the moment of nth arrival τ_n. Let the overflow demand, $\mathscr{O}_{n,k}^{(m)}$, denote indices of the forwarded demand m, active at time τ_n, that failed to be met by 'safety stock' impressions of capacity ΔC_k and ended up using a subset of impressions from the shared pool of capacity $\hat{C}_k$. Also, let $J_n(m,k)$ be an indicator function equal to 1 if nth transaction is of type m; otherwise, $J_n(m,k) = 0$. Then, given that the amount of delivered impressions can not be larger than $A \triangleq B/(\min_k p_k)$, one can upper bound $p_{m,k}^{(b)}(C_k, \Delta C_k)$ as

$$\begin{aligned} & p_{m,k}^{(b)}(C_k, \Delta C_k) = \\ & \mathbb{P}[\sum_{m=1}^{M} \sum_{i \in \mathscr{N}_{n,k}^{(m)}} 1[\alpha_k^{(m)} > 0] B_i^{(m)} / p_k \\ & \quad + \sum_{m=1}^{M} \sum_{i \in \mathscr{O}_{n,k}^{(m)}} B_i^{(m)} / p_k + B_n^{(m)} / p_k > C_k - \Delta C_k, J_n(m,k) = 1] \\ & \leq \mathbb{P}[\sum_{m=1}^{M} \sum_{i \in \mathscr{N}_{n,k}^{(m)}} 1[\alpha_k^{(m)} > 0] B_i^{(m)} / p_k \\ & \quad + \sum_{m=1}^{M} \sum_{i \in \mathscr{O}_{n,k}^{(m)}} B_i^{(m)} / p_k > C_k - \Delta C_k - A]. \end{aligned} \tag{27, 28}$$

Next, we concentrate on estimating the overflow traffic $\mathscr{O}_{n,k} \triangleq \cup_{m=1}^{M} \mathscr{O}_{n,k}^{(m)}$. Note that the set of the overflow transactions at time τ_n can be enlarged if all of the transaction demand that can not be met by the LP suggested, initial, choice of inventory, is redirected to inventory k. Therefore,

$$\begin{aligned} & \mathscr{O}_{n,k} \subset \\ & \cup_{i<n} \{i : \sum_{l \neq k} \sum_{m=1}^{M} \sum_{j \in \mathscr{N}_{i,l}^{(m),*}} 1[\alpha_l^{(m)} > 0] B_j^{(m)} / p_l \\ & \quad > \sum_{l \neq k} [C_l - \Delta C_l] + \Delta C_k / A - KA, \cap_m D_i^{(m)} > \tau_n - \tau_i\}, \end{aligned} \tag{29}$$

where we use $\mathscr{N}_{i,l}^{(m),*}$ to denote a set of indices of transactions which original choice of inventory is $l \neq k$, that arrived before τ_i and are still active at the moment of ith arrival τ_i.

Expression (29), accumulates all of the excess demand that is forwarded from the original LP-based inventory choice and is active at τ_n. Since the same budget translates into different amounts of impressions depending on their price, we further increase the overflow set by dividing the safety stock capacity ΔC_k by the maximum possible resource requirement $A = B/\min_k p_k$, since

$$\frac{B_j^{(m)}}{p_k} = \frac{B_j^{(m)}}{p_l} \frac{p_l}{p_k} \leq \frac{B_j^{(m)}}{p_k} A;$$

The previous expression addresses the amount of impressions consumed by forwarded transactions. For example, if a transaction with budget $B_j^{(m)}$ is forwarded from resource l to resource k, its demand of $B_j^{(m)}/p_l$ impressions translates into the demand of $B_j^{(m)}/p_k$ impressions, and the ratio of these two quantities is at most A. Furthermore, when we define superset in (29), we deduct KA from the right hand side in order to account for the potentially positive leftover capacity in inventories $l \neq k$ that is too small to meet the incoming demand.

The key observation is that the sum of the transaction demand in (29) corresponds to the overflow 'traffic' in the uncapacitated M/G/∞ system. Given that the resource requirements take values between 1 and A, and that the budget values are mutually independent, we can rewrite the sum in (29), by regrouping the independent Poisson streams of transactions based on the corresponding impression requirements, i.e.,

$$\sum_{l \neq k} \sum_{m=1}^{M} \sum_{j \in \mathscr{N}_{i,l}^{(m),*}} 1[\alpha_l^{(m)} > 0] B_j^{(m)} / p_l = \sum_{s=1}^{A} s Y_i^{(s)}, \tag{30}$$

where

$$Y_i^{(s)} \triangleq \sum_{m=1}^{M} \sum_{l \neq k} \sum_{j \in \mathscr{N}_{i,l}^{(m)*}} 1[\alpha_l^{(m)} > 0] 1[B_j^{(m)} = p_l s].$$

Next, we further enlarge $\mathscr{O}_{n,k}$ as

$$\mathscr{O}_{n,k} \subset \cup_{i<n} \cup_{s=1}^{A} \{i : s Y_i^{(s)} > s \mathbb{E} Y_i^{(s)} + \Delta C_k / A^2, D_i > \tau_n - \tau_i\}, \tag{31}$$

where we use D_i to denote the duration of the ith request for s impressions. Finally, using the result of Theorem 2, we show that for

$$\Delta C_k = \kappa\sqrt{C\log C}, \tag{32}$$

where κ is a constant that depends on the system parameters such as the maximum impression requirement A, we obtain that for $C \to \infty$ large enough processes

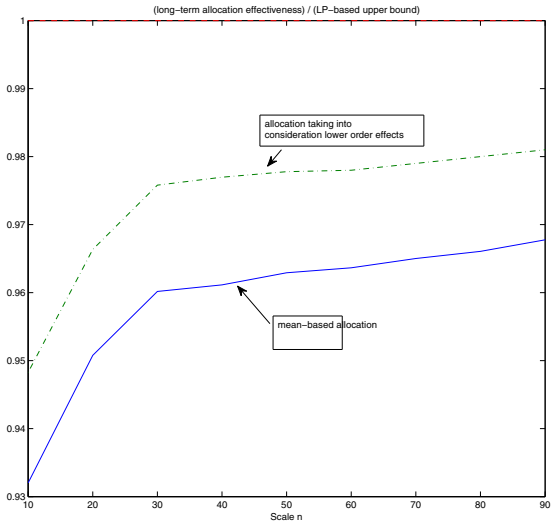
$$\mathscr{Y}_n^{(s)} \triangleq \cup_{i<n}\{i : sY_i^{(s)} > s\mathbb{E}Y_i^{(s)} + \Delta C_k/A^2 - K, D_i > \tau_n - \tau_i\},$$

are bounded by independent Poisson processes with finite rate. This, in conjunction with (31), (28) and $B_i^{(m)} < B < \infty$, allows us to finally apply generalized Erlang formula and show that $p_{m,k}^{(b)}(C_k, \Delta C_k)$ converges to zero as $\rho \to \infty$, which completes the outline of the main proof. $\diamondsuit$

VII. Numerical Example

In this section we present a numerical example capturing the phenomena we tried to analyze in this paper. We assume that there are three inventory types: expensive ($ 10 for 1000 impressions), moderately expensive ($ 5 for 1000 impressions) and cheap ($1 per 1000 impressions). Inventory's price is in correlation with its effectiveness measure, and in this case we have $\varepsilon_1 = 0.2$, $\varepsilon_2 = 0.05$ and $\varepsilon_3 = 0.0001$ for the expensive, moderately expensive and remnant inventory. Furthermore, we assume that the most expensive inventory has the smallest capacity, while the less expensive is usually more available ($C_1 = 5 \times 10^6$, $C_2 = 2C_1$ and C_3 is unlimited. Similarly, we segment advertisers by their budget into three basic groups: high, moderate and low budget ones. We assume that most of the advertisers have moderate budget.

The figure depicts an example of system's performance, i.e., proportion of the LP-based net revenue upper bound achieved for different scales r, in the case where inventory management is done based on mean-value allocation, as well as when lower order terms originating from forwarded demand are considered. Here, we assume that $\gamma^{(1)} = \gamma^{(2)} = \gamma^{(3)} = 1$. We see that the system can achieve significant performance benefits when allocation policy consideres variations in the amount of forwarded traffic, i.e., provides safety stocks.



References

[1] M. Abramowitz and I. A. Stegun, editors. *Handbook of Mathematical Functions*. Dover Publications, Inc., New York, 1974.

[2] S. M. Berman. Limit theorems for the maximum term in stationary sequences. *Annals of Mathematical Statistics*, 35:502–516, 1964.

[3] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004.

[4] N. G. Duffield and W. Whitt. Control and recovery from rare cingestion events in a large multi-server system. *Queueing Systems*, 26:69–104, 1997.

[5] W. N. Kang, F. Kelly, N. H. Lee, and R. J. Williams. State space collapse and diffusion approximation for a network operating under a fair bandwidth-sharing policy. *Annals of Applied Probability*, 19:1719–1780, 2009.

[6] F. Kelly. Blocking probabilities in large circuit-switched networks. *Advances in Applied Probability*, 18(2):473–505, 1986.

[7] F. Kelly. Loss networks. *Annals of Applied Probability*, 1(3):319–378, 1991.

[8] M. R. Leadbetter. On extreme values in stationary sequences. *Probability Theory and Related Fields*, 28(4):289–303, 1974.

[9] R. Levi and A. Radovanović. Provably near-optimal LP-based policies for revenue management in systems with reusable resources. Technical Report TR 4702-08, Massachusetts Institute of Technology, Sloan School of Management, Cambridge, 2007.

[10] R. Levi and A. Radovanovic. Provably near-optimal lp-based policies for revenue management in systems with reusable resources. *Operations Research*, 58(2):503–507, 2010.

[11] Y. Lu and A. Radovanović. Asymptotic blocking probabilities in loss networks with subexponential demands. *http://arxiv.org/abs/0708.4059*, September 2007.

[12] Y. Lu and A. Radovanović. Asymptotic blocking probabilities in loss networks with subexponential demands. *Journal of Applied Probability*, 44(4):1088–1102, 2007.

[13] Y. Lu, A. Radovanović, and M. S. Squillante. Optimal capacity planning in stochastic loss networks. In *SIGMETRICS Performance Evaluation Review*, volume 25, 2007.

[14] A. Papoulis. *Probability, random variables and stochastic processes*. McGraw-Hill Book Company, 1965.

[15] B.A. Sevastyanov. An ergodic theorem for Markov processes and its application to telephone systems with refusals. *Theory of probability and its applications*, 2:104–112, 1957.

[16] D. Slepian. The one-sided barrier problem for gaussian noise. *Bell System Tech. J.*, 41:463–501, 1962.

[17] W. Whitt. Blocking when service is required from several facilities simultaneously. *AT&T Bell Lab. Tech. Journal*, 64(8):1807–1856, 1985.