Information Exchange Surrogates for Approximation of Blocking Probabilities in Overflow Loss Systems

Eric W. M. Wong*, Jun Guo*, Bill Moran† and Moshe Zukerman*
*Department of Electronic Engineering, City University of Hong Kong, Hong Kong SAR, China
Email: {eeewong, j.guo, m.zu}@cityu.edu.hk
†Defence Science Institute, The University of Melbourne, Australia
Email: wmoran@unimelb.edu.au

***Abstract*—Overflow loss systems are an important class of teletraffic models. Evaluation of blocking probabilities in such systems involving mutual overflow effects is a difficult problem. In the literature, decoupling a given system into independent subsystems is typically regarded as a scalable, though non-robust, approach to the problem. This paper presents a new method that is based on a radically different idea from that of the conventional approach. Firstly a surrogate model that, in a systematic way, approximately captures the state dependencies due to the overflow model is designed. Secondly it is observed that approximation of the blocking probability of the surrogate model provides a good approximation to the blocking probability in the original model. We introduce important concepts underpinning this surrogate-based approximation method, and demonstrate its effectiveness by applying it to an overflow model that incorporates mutual overflow effects common to various applications of overflow loss systems. Extensive and statistically reliable experiments demonstrate that the new method yields significantly and consistently better results compared to the conventional approach, improving the accuracy by orders of magnitude in many instances and yet requiring less computational effort.**

I. Introduction

We consider an important class of teletraffic models called *overflow loss systems*. They arise in a variety of contexts of telecommunication systems, including gradings [1], [2], circuit-switched networks using alternate routing [3], [4], and optical burst-switching networks [5], [6]. They have also found emerging applications in service sectors for modeling, e.g., video server systems [7], [8], call centers [9], [10], health-care systems [11], [12], and have led to potential solutions to challenging problems for cost-effective operations of such systems subject to stringent quality of service requirements.

In general, overflow loss systems are characterized by calls/requests requiring service in a system comprising multiple server groups. An incoming call either is admitted to a server group with at least one idle server or overflows to another server group. If all server groups accessible to the call are busy, it is blocked and cleared from the system. The probability that calls are blocked and cleared from the system, known as the *blocking probability*, is an important performance measure of overflow loss systems.

It is well known in traffic engineering that evaluation of blocking probabilities in overflow loss systems is a difficult problem [13]. This is particularly true for non-hierarchical models where overflow from any server group may directly or indirectly affect the load of any other server group. In particular, the *mutual overflow* effect [3] refers to a situation where there is congestion on a specific server group causing overflow to the other server groups, and where this overflow loads up the other server groups so that they in turn yield overflow back to the original server group. Such models, in many practical cases, are not amenable to an exact analysis because they exhibit significant state dependencies, which make the state-space required for an exact analysis too large [14]. The challenge is to find a robust methodology for approximations of such models to capture their overflow-induced state dependencies in a scalable way.

This paper proposes a new and versatile approach, called *information exchange surrogate approximation* (IESA), for development of robust approximation algorithms for blocking probability evaluation in non-hierarchical overflow loss systems. The key philosophy that guides the design of IESA is to establish a certain *surrogate model* that in a systematic way approximately captures the state dependencies due to mutual overflow in the original model. It is expected that the surrogate model yields a close but somewhat different blocking probability from that of the original model such that the error introduced in approximation of the surrogate model ideally cancels out the difference between the blocking probability of the surrogate model and that of the original model. We call such a surrogate-based approach a *surrogate approximation*.

IESA is based on a type of surrogate models called an *information exchange system* (IES). An incoming call in an IES that finds a server group busy may exchange certain congestion information with a call in service at the server group before overflowing to other server groups. Such an imbedded *information exchange mechanism* allows overflowing calls to propagate congestion information in the system and also dictates selections of server groups for subsequent attempts by an overflowing call. An IES is different from the original model where there is no such information exchange mechanism and the overflow decision of a call is independent of past experience of other calls. IESA aims to find a "right" IES such that approximation of the IES yields a robust and scalable approximation of the original model.

IESA can be applied to any application of overflow loss systems. In this paper, we demonstrate the effectiveness of IESA by applying it to a model that incorporates mutual overflow effects commonly seen in various applications of overflow loss systems. This model makes it convenient to introduce important concepts underpinning IESA and suitable for exposing the weaknesses of conventional approaches proposed

in the literature. We observe that an IESA algorithm based on an appropriate IES can lead to significant improvement in accuracy compared to the conventional approaches.

The rest of the paper is organized as follows. In Section II, we review the conventional approaches proposed in the literature for blocking probability evaluation in non-hierarchical overflow loss systems. Section III provides the model of the overflow loss system considered in this paper. In Section IV, we describe the set of call attributes used in defining an IES and introduce IES1 and IES2 as two IES-type surrogate models. Section V presents detailed equations that form the two IES-based surrogate approximation algorithms. Numerical results are provided in Section VI to demonstrate the effectiveness of IESA. Section VII draws conclusions.

II. Related Work

McNamara [14] showed that non-hierarchical overflow loss systems do not have product-form solutions for the blocking probability. The problem, with the assumptions of Poisson arrivals and exponentially distributed service times, can only be solved exactly by a multi-dimensional Markov process. Although the blocking probability can, in principle, be obtained by solving numerically a set of steady-state equations, this approach is not scalable because of the curse of dimensionality. For the simplest case where the system comprises only two server groups, McNamara managed to compute the exact solutions only for a symmetric system with up to five servers in each server group.

Koole and Talim [9] modeled multi-skill call centers as non-hierarchical overflow loss systems and proposed a method called *exponential approximation* (EA) for evaluation of blocking probabilities. EA is equivalent to a classic method first introduced in [15] and now widely known as the *Erlang fixed point approximation* (EFPA) [16] for evaluation of blocking probabilities in circuit-switched networks. Guo et al. [8] modeled video server systems as non-hierarchical overflow loss systems and applied a similar approach to EA for evaluation of blocking probabilities.

EA is based on decoupling the given system into independent server groups (subsystems), and assuming that the aggregate of original traffic and overflow traffic offered to a server group of N servers follows a Poisson process. In this way, it is able to treat each server group as an independent $M/M/N/N$ system, and the probability that a server group is busy is approximated by the Erlang B formula [17]. This approach inherently gives rise to non-linear equations that entail a fixed-point solution of the blocking probability.

EA relies on two fundamental assumptions that are rarely completely satisfied:

- **Poisson assumption**. EA assumes that the overflow traffic follows a Poisson process, whereas it is known to have higher variance than a Poisson process [18].
- **Independence assumption**. EA assumes that server groups are mutually independent, whereas they are in fact statistically dependent because a large number of busy servers in a server group may indicate a heavy traffic period, so that other server groups are also likely to be heavily loaded at that time.

The effect of the Poisson assumption can be mitigated by characterizing overflow traffic using moment-matching techniques [18], [19]. However, it was observed in [20] that the reduction in error is marginal, implying that the dominant source of error in such systems is the independence assumption. Holtzman [21] proposed a method of taking dependence into account by approximating conditional probabilities which reflect the dependence. However, such a brute-force approach directly applied to the given system is not scalable.

The *overflow priority classification approximation* (OPCA) proposed in [20] was the first notable surrogate approximation algorithm for addressing the difficult problem of blocking probability evaluation in non-hierarchical overflow loss systems. OPCA is based on a surrogate model called in this paper a *preemptive priority system* (PPS). PPS is constructed by introducing a *preemptive priority regime* to calls in the original model in such a way that each call is classified according to its *seniority* in terms of the number of times it overflows. Classifying calls according to their seniority creates a multi-level traffic hierarchy in PPS where the n-th level of the hierarchy includes calls of seniority n or lower. In this way, OPCA essentially transforms the original model that is a non-hierarchical system into a surrogate model that is a hierarchical system. Applying decomposition in each level of the analysis for the hierarchical surrogate model provides a computationally efficient way for estimating its blocking probability. In particular, for overflow loss systems where a call requires service from only one server group, the approximation can be obtained hierarchically in a finite number of steps without the need for a fixed-point solution.

As demonstrated in [20] and further demonstrated in this paper, OPCA is a scalable surrogate approximation algorithm that is accurate for systems where calls have full or almost full accessibility to server groups. It is a rudimentary approach to modeling state dependencies and requires substantial modifications and adaptations to make it versatile to deal with generic limited-accessibility systems. The new IESA approach proposed here has its historical roots in OPCA. However, they differ fundamentally at a conceptual level. The core discriminating idea of IESA from OPCA is the conceptual replacement of the preemptive priority regime supporting OPCA by the information exchange mechanism underpinning IESA.

III. The Model

Consider an overflow loss system with M traffic sources forming the set $\mathcal{F}$ and k server groups forming the set $\mathcal{D}$. Calls/requests from source $m \in \mathcal{F}$ have access to n_m different server groups; the set of these server groups is denoted by $\Gamma_m \subseteq \mathcal{D}$. The set of traffic sources that have access to server group $d \in \mathcal{D}$ is denoted by $\Phi_d \subseteq \mathcal{F}$. Let $k_d^* = \max_{m \in \Phi_d} n_m$, $d \in \mathcal{D}$. The number of servers on server group $d \in \mathcal{D}$ is N_d.

When a request from source m arrives, the system randomly selects with equal probability a server group $d \in \Gamma_m$. If the number of calls being served by server group d is less than N_d, the request is admitted; otherwise, the request randomly attempts (with equal probability among the remaining server groups) another server group in Γ_m. This process continues until either the request is served by one of the server groups in Γ_m, or all server groups in Γ_m are found to be busy. In the latter case, it is blocked and cleared from the system.

A request from source m is defined as a 0-call if it has just been initiated. The request becomes an n-call, $1 \leq n \leq n_m - 1$, if it has overflowed n times. When the request becomes an n_m-call, it is blocked and cleared from the system.

Requests from source m form an independent Poisson arrival process with rate λ_m. The service time of a request from source m is exponentially distributed with mean $1/\mu_m$. The offered traffic from source m is $A_m = \lambda_m/\mu_m$. The offered traffic to the system is $A = \sum_{m \in \mathcal{F}} A_m$.

IV. Information Exchange System

This section provides the details of our design of the IESA approach. We begin by describing the set of call attributes used in defining an IES. Then, we introduce IES1 as an equivalent surrogate model to PPS that provides an information exchange viewpoint on PPS. We discuss how the insight gained from the success and limitation of OPCA motivates the design of IES2 as a more appropriate surrogate model in generic limited-accessibility systems.

A. Call attributes

Calls in IES are designed to have several attributes. The first is the call identity, denoted by I, that includes the traffic source where the call is initiated and the elapsed time since the start of its service. The second is a set of server groups, denoted by Δ, that the call has already attempted and overflowed from. The specific choice of the other attributes is part of the definition of a particular IES. They are normally based on information received from other calls. In this paper, we consider IES where calls have not more than three attributes, but in principle, the definition of IES can be further extended to include additional attributes. In the case of an IES with three attributes, the third attribute is denoted by Ω.

Let I_i, Δ_i and Ω_i denote the first, second and third attribute (if it exists) of request i. Henceforth, we will use (I, Δ)-call and (I, Δ, Ω)-call to denote calls in IES with two and three attributes, respectively. This notation is more comprehensive than the n-call notation used in the original model where a call is simply characterized by the number of times it has overflowed. However, in contexts where the number of overflows so far is sufficient to characterize calls in IES, we will still use the shorter n-call notation.

In general, some or all of the attributes of a given call may change during the sojourn time of the call in the system and will dictate when the call leaves the system. The attributes of a call in IES may change when it meets other calls in the system and exchange certain information. The particular rule by which these attributes change is also a part of the definition of the particular IES under consideration. We use the information exchange features of IES to develop new surrogate models that give rise to new approximation algorithms for blocking probability evaluation in non-hierarchical overflow loss systems.

B. IES1

As an illustration of the power of IES we show how, as a special case, it can encompass PPS. IES1 is constructed to be equivalent to PPS but provides an information exchange viewpoint on PPS. Let calls in both systems be described using the (I, Δ)-call notation. Let $|\Delta|$ be the cardinality of the set Δ. An (I_1, Δ_1)-call that attempts an unavailable server group busy serving an (I_2, Δ_2)-call, $|\Delta_1| < |\Delta_2|$, is a *junior call*, while the (I_2, Δ_2)-call in service is a *senior call*. In PPS, the junior call is given the right to preempt the senior call and access the server group. The preempted call will then overflow and attempt server groups that it has not attempted before. If a server group is available, the service period of the preempted call continues from where it was last preempted. On the other hand, instead of preempting the senior call, the junior call in IES1 exchanges its two attributes with those of the senior call (including seniority, i.e., $|\Delta|$) and then overflows. Thus, if PPS and IES1 are both loaded with exactly the same arrival process, at any point in time, for any call in either PPS or IES1, there exists a corresponding call in the other system that has the same attributes and is in the same position (either in service or being overflowed). This one-to-one correspondence makes PPS and IES1 equivalent, so they have the same blocking probability.

In its general form, IES1 is described as follows. Consider an (I_1, Δ_1)-call from source m with $|\Delta_1| < n_m$ that arrives at server group $d \in \Gamma_m$. If the server group has one or more servers idle, the call is admitted. Otherwise, if the most senior call that the server group is serving is an (I_2, Δ_2)-call with $|\Delta_1| \geq |\Delta_2|$, the arriving call overflows and becomes an $(I_1, \Delta_1 \cup \{d\})$-call. However, if $|\Delta_1| < |\Delta_2|$, we have a case where a junior call attempts a server group serving a senior call. In such a case, the two calls exchange their attributes. In particular, the call in service becomes an (I_1, Δ_1)-call while the arriving call overflows and becomes an $(I_2, \Delta_2 \cup \{d\})$-call. In either case, if the cardinality of the second attribute of the overflowed call reaches n_m given that it is a request from source m, the call is blocked and cleared from the system.

An overflowed (I, Δ)-call from source m that is not blocked will never attempt server groups that are included in Δ. It will attempt with equal probability other server groups in Γ_m. A new call from source m has Δ set to $\emptyset$ (empty set) and can try with equal probability any server group in Γ_m. Thus, in IES1, the cardinality of the second attribute of any call that overflows from server group d is at most k_d^*.

IES1 is equivalent to PPS and therefore does not lead directly to a more accurate approximation than OPCA. Nevertheless, IES1 is important for the following two reasons:

- IES1 provides a convenient approach to understanding why OPCA provides more accurate blocking probability evaluation in full-accessibility systems than EA.
- IES1 is based on an information exchange mechanism. The insight gained from its success and limitation serves as a guideline to the design of more robust approximation algorithms under the IESA framework.

IES1 by nature has a higher blocking probability than that of the original model. To explain this effect, consider a junior call in IES1 that attempts an unavailable server group busy serving a senior call. The junior call will exchange its two attributes with the senior call and then overflow. In this way, the overflowed call "forgets" the information of unavailable server groups carried in its own second attribute. Instead, it will

subsequently use the information of unavailable server groups previously carried by the senior call, and hence will not attempt those server groups. There is, however, a positive probability that the congestion information is outdated. That is, one or more of those server groups presumed to be unavailable might have become available during the sojourn time of the senior call in the system. The rule defined by IES1 does not allow the overflowed call to access such available server groups. In contrast, an overflowed call in the original model is allowed to access any server group that is available. As a result, IES1 has a higher blocking probability compared to the original model.

Since the information exchange mechanism allows IES1 to gather the congestion information of server groups in the system, as we build the multi-level traffic hierarchy for approximation, more and more state dependencies can be captured and embedded in the overflow traffic in each subsequent level of the analysis. In each level of the analysis, we make the two fundamental assumptions of EA discussed in Section II. As a result, the approximation of IES1 underestimates its blocking probability. Nevertheless, the impact of such simplifying assumptions on the approximation error is likely to be weaker than in the case of EA. This is because with EA there is no specific mechanism to capture the state dependencies. Therefore, one can expect that the error introduced by approximation of IES1 is likely to be smaller than that introduced by EA to the original model. Given the fact that the blocking probability of IES1 is greater than that of the original model, one can further expect that the blocking probability predicted by OPCA will be greater than that of EA.

As demonstrated in [20] and further demonstrated in Section VI of this paper, for systems where calls have full or almost full accessibility to server groups, OPCA is very accurate which implies that IES1 is the right surrogate model. However, for systems where calls have more limited accessibility to server groups, IES1 can be very different from the original model which leads to inaccurate approximations.

To illustrate this effect, let us consider the system where each of the k server groups has only one server. Consider a 0-call from source 1 that has access to server groups 1 and 2. Assume that server group 1 is busy serving a 1-call which is a request from source 2 that has access to server groups 1 and 3. Also, assume that server group 2 is available. The new call from source 1 attempts server group 1 and is rejected because of the 1-call in service. The senior call has already attempted server group 3 and has been rejected there. Under IES1, the junior call exchanges both attributes with the senior call, and will be blocked and cleared from the system because it becomes a 2-call from source 2. Note that, before information exchange, the junior call should have been given the opportunity to attempt server group 2 where a call from source 1 has access. The information that server group 3 is unavailable was irrelevant for the junior call when it was a 0-call from source 1. The outcome here is a loss of a request, which would not have happened in the original model.

In systems where calls have more limited accessibility to server groups, the probability of a call in IES1 "paying" for this kind of irrelevant information rises, as a result of which, OPCA increasingly overestimates the blocking probability of the original model. On the other hand, in the case of full-accessibility systems where OPCA is known to be very accurate, since $n_m = k$, $m \in \mathcal{F}$, this effect never happens.

C. IES2

The success and limitation of IES1 motivates us to design IES2 as a more appropriate surrogate model in generic limited-accessibility systems. Calls in IES2 have a third attribute Ω representing an estimate of the number of unavailable server groups in the system. Unlike IES1, the seniority of a call in IES2 is determined by the value of Ω. The information exchange in IES2 between a junior call and a senior call involves only the third attribute and never involves the first and second attributes. In this way, IES2 behaves more like the original model, where an overflowed call retains its call identity (using the first attribute) and its actual overflow record (using the second attribute) during its hunt for an available server group. Meanwhile, exchange of the third attribute allows the overflowed call in IES2 to gather the congestion information of other server groups to capture the state dependencies in the original model and ensures a higher blocking probability in IES2 than in the original model.

Formally, IES2 is described as follows. A new call entering the system has Δ set to $\emptyset$ and Ω set to 0. Consider an $(I_1, \Delta_1, \Omega_1)$-call from source m with $|\Delta_1| < n_m$ and $\Omega_1 < k$ that arrives at server group $d \in \Gamma_m$. If the server group has one or more servers idle, the call will access the server group. Otherwise, if the most senior call that the server group is serving is an $(I_2, \Delta_2, \Omega_2)$-call with $\Omega_1 \geq \Omega_2$, the arriving call overflows and becomes an $(I_1, \Delta_1 \cup \{d\}, \Omega_1 + 1)$-call. However, if $\Omega_1 < \Omega_2$, we have a case where a junior call attempts a server group serving a senior call. In such a case, the two calls exchange their third attribute. In particular, the call in service becomes an $(I_2, \Delta_2, \Omega_1)$-call while the arriving call overflows and becomes an $(I_1, \Delta_1 \cup \{d\}, \Omega_2 + 1)$-call. In this way, for an arriving (I, Δ, Ω)-call in IES2, we will always have $|\Delta| \leq \Omega$. This implies that the number of server groups that the request has already attempted (and overflowed) is a lower bound of the estimate of the number of server groups that are unavailable in the entire system.

An overflowed (I, Δ, Ω)-call from source m that is not blocked will never attempt server groups that are included in Δ. For the remaining $n_m - |\Delta|$ server groups in Γ_m, there is a certain probability P that they are all unavailable. To evaluate this probability, the (I, Δ, Ω)-call from source m assumes that Δ is included in the set of server groups presumed unavailable, and evaluates P as follows:

$$P = \begin{cases} 0, \text{ if } \Omega < n_m; \\ \dfrac{\binom{\Omega - |\Delta|}{n_m - |\Delta|}}{\binom{k - |\Delta|}{n_m - |\Delta|}}, \text{ if } \Omega \geq n_m. \end{cases} \quad (1)$$

Notice that an (I, Δ, Ω)-call from source m that is still in the system must satisfy $|\Delta| \leq \Omega < k$ and $|\Delta| < n_m$, so we have $0 \leq P < 1$.

With probability P, the (I, Δ, Ω)-call is blocked and cleared from the system. With probability $1 - P$, it will instead continue to attempt a server group in $\Gamma_m - \Delta$ with probability $1/(n_m - |\Delta|)$. Then, the process repeats itself until the call is admitted, $|\Delta|$ reaches n_m, or Ω reaches k; in either of the two latter cases, the call is blocked and cleared from the system.

V. APPROXIMATION

This section presents the detailed form of the EA and two IES-based surrogate approximation algorithms. We begin by showing how to apply EA to the original model. Then, we derive equations that represent the IES1-based surrogate approximation algorithm, which we call IESA1. In particular, we describe the multi-level traffic hierarchy embedded in IES1. This provides a fundamental hierarchy based on information exchange, and hence a steppingstone for development of the more robust IES2-based surrogate approximation algorithm, which we call IESA2.

The nature of random hunting for an available server group requires account to be taken of each random sequence of server groups that a request attempts. For this purpose, we define $\Psi(X, x)$, $x = 0, 1, \ldots, |X|$, as the set of ordered choices of x elements from X. By definition, $\Psi(X, 0) = \emptyset$.

A. EA

To apply EA to the original model, the system of k server groups is decoupled into k independent subsystems, where the d-th subsystem is treated as an $M/M/N_d/N_d$ system having load equal to the original traffic plus all the traffic that overflows to it from the other server groups.

For each $d \in \mathcal{D}$ in the original model, define

- $a_{d,m,n,\boldsymbol{s}}$ – Offered traffic to server group d ($d \notin \boldsymbol{s}$) made up of n-calls from source m that have overflowed sequentially from n server groups in Γ_m along the path $\boldsymbol{s} = \{s_1, s_2, \ldots, s_n\}$, $m \in \Phi_d$, $n = 0, 1, \ldots, n_m - 1$.
- $a_{d,n}$ – Offered traffic to server group d made up of n-calls, $n = 0, 1, \ldots, k_d^* - 1$.
- A_d – Combined traffic offered to server group d made up of 0-calls, 1-calls, …, or $(k_d^* - 1)$-calls; namely,
$$A_d = \sum_{n=0}^{k_d^*-1} a_{d,n}.$$
- $v_{d,m,n,\boldsymbol{s}}$ – Overflow traffic from server group d ($d \in \boldsymbol{s}$) made up of n-calls from source m that have overflowed sequentially from n server groups in Γ_m along the path $\boldsymbol{s}$ ($s_n = d$), $m \in \Phi_d$, $n = 1, 2, \ldots, n_m$.
- B_d – Probability that all servers of server group d are busy, serving n-calls with $n = 0, 1, \ldots,$ or $k_d^* - 1$.

Summing $a_{d,m,n,\boldsymbol{s}}$ over all eligible $m \in \Phi_d$ and each n-element path $\boldsymbol{s}$ from the set $\Psi(\Gamma_m - \{d\}, n)$, we obtain

$$a_{d,n} = \sum_{m \in \Phi_d, n_m > n} \sum_{\boldsymbol{s} \in \Psi(\Gamma_m - \{d\}, n)} a_{d,m,n,\boldsymbol{s}} \tag{2}$$

for $n = 0, 1, \ldots, k_d^* - 1$, where $a_{d,m,0,\emptyset} = A_m / n_m$.

With the independence assumption, the offered traffic $a_{s_n,m,n-1,\boldsymbol{s}-\{s_n\}}$ to server group s_n, which has overflowed sequentially from $n-1$ server groups along the path $\boldsymbol{s}$, is overflowed with probability B_{s_n} from server group s_n and then becomes the overflow traffic $v_{s_n,m,n,\boldsymbol{s}}$. With probability $1/(n_m - n)$, the overflow traffic $v_{s_n,m,n,\boldsymbol{s}}$ will be offered to server group d, given that it is not in the path $\boldsymbol{s}$ and therefore has not been attempted. Accordingly, we have

$$a_{d,m,n,\boldsymbol{s}} = \frac{v_{s_n,m,n,\boldsymbol{s}}}{n_m - n} \tag{3}$$

and $v_{s_n,m,n,\boldsymbol{s}} = a_{s_n,m,n-1,\boldsymbol{s}-\{s_n\}} B_{s_n}$. Thus, we can derive $a_{d,m,n,\boldsymbol{s}}$ as

$$a_{d,m,n,\boldsymbol{s}} = \frac{A_m}{n_m} \prod_{i=1}^{n} \frac{B_{s_i}}{n_m - i} \tag{4}$$

and A_d as

$$A_d = \sum_{m \in \Phi_d} \frac{A_m}{n_m} \left[1 + \sum_{n=1}^{n_m - 1} \sum_{\boldsymbol{s} \in \Psi(\Gamma_m - \{d\}, n)} \prod_{i=1}^{n} \frac{B_{s_i}}{n_m - i} \right]. \tag{5}$$

Using the Poisson assumption, we calculate the probability B_d using the Erlang B formula

$$B_d = \mathbf{E}(A_d, N_d). \tag{6}$$

For compatibility, EA requires that the B_d values so calculated be the same as those used to calculate the reduced load in (5). Thus, (5) and (6) constitute a set of fixed-point equations, which can often be solved by a successive substitution method. In our numerical examples, we use the method of [22] for solving the B_d values. The iteration is continued until the changes of the B_d values are less than 10^{-8}.

It follows that the overall blocking probability of the original model predicted by EA is given by

$$B_{\mathrm{EA}} = 1 - \frac{\sum_{d \in \mathcal{D}} A_d (1 - B_d)}{A}. \tag{7}$$

In (7), $1 - B_d$ is the probability that server group d has one or more servers idle. Therefore, $A_d(1 - B_d)$ is the total traffic carried by server group d estimated by EA. The blocking probability of calls from source m is given by

$$\hat{B}_m = \frac{\sum_{\boldsymbol{s} \in \Psi(\Gamma_m, n_m)} v_{s_{n_m},m,n_m,\boldsymbol{s}}}{A_m}. \tag{8}$$

B. IESA1

In IES1, classifying calls according to their seniority creates a hierarchy where the n-th level of the hierarchy includes calls of seniority n or lower. The seniority of a call in service in server group $d \in \mathcal{D}$ of IES1 is at most $k_d^* - 1$.

IESA1 is based on the following multi-level traffic hierarchy considered for each server group in IES1 to capture the exchange of the two attributes. At level 0, the offered traffic to server group d is made up of new calls (with $|\Delta| = 0$) only. An arriving call that finds the server group available is admitted; otherwise, the call is rejected. At each subsequent level $n = 1, 2, \ldots, k_d^* - 1$, the offered traffic to server group d is made up of the offered traffic to it at level $n-1$ plus all calls that overflow at level $n-1$ from other server groups and hence are offered as calls with $|\Delta| = n$ to server group d. An arriving call, with $|\Delta| = 0, 1, \ldots,$ or n, that finds the server group available is admitted; otherwise, all servers of the server group are busy, serving calls with $|\Delta| = 0, 1, \ldots,$ or n. If the seniority of the arriving call is lower than that of the most senior call in service, the arriving call overflows

with information exchange; otherwise, it overflows without information exchange.

In each level of the analysis, IESA1 assumes that the offered traffic to each server group in IES1 is Poisson and the states of different server groups are mutually independent. Thus, the system of k server groups is decoupled into k independent subsystems, where the d-th subsystem is treated as an $M/M/N_d/N_d$ system.

For each $d \in \mathcal{D}$ in the surrogate IES1 model, define

- $a^{\text{IES1}}_{d,m,n,\boldsymbol{s}}$ – Offered traffic to server group d ($d \notin \boldsymbol{s}$) made up of calls from source m with $|\Delta| = n$ that have overflowed sequentially from n server groups in Γ_m along the path $\boldsymbol{s} = \{s_1, s_2, \ldots, s_n\}$, $m \in \Phi_d$, $n = 0, 1, \ldots, n_m - 1$.
- $a^{\text{IES1}}_{d,n}$ – Offered traffic to server group d made up of calls with $|\Delta| = n$, $n = 0, 1, \ldots, k^*_d - 1$.
- $A^{\text{IES1}}_{d,n}$ – Offered traffic to server group d at level n, $n = 0, 1, \ldots, k^*_d - 1$.
- $v^{\text{IES1}}_{d,m,n,\boldsymbol{s}}$ – Overflow traffic from server group d ($d \in \boldsymbol{s}$) made up of calls from source m with $|\Delta| = n$ that have overflowed sequentially from n server groups in Γ_m along the path $\boldsymbol{s}$ ($s_n = d$), $m \in \Phi_d$, $n = 1, 2, \ldots, n_m$.
- $v^{\text{IES1}}_{d,n}$ – Overflow traffic from server group d made up of calls with $|\Delta| = n$, $n = 1, 2, \ldots, k^*_d$.
- $b^{\text{IES1}}_{d,n}$ – Probability that a call with $|\Delta| = n$ overflows from server group d at level n, $n = 0, 1, \ldots, k^*_d - 1$.
- $B^{\text{IES1}}_{d,n}$ – Probability that all servers of server group d at level n are busy, serving calls with $|\Delta| = 0, 1, \ldots,$ or n, $n = 0, 1, \ldots, k^*_d - 1$.

By definition, we have $A^{\text{IES1}}_{d,n} = A^{\text{IES1}}_{d,n-1} + a^{\text{IES1}}_{d,n}$, $n = 0, 1, \ldots, k^*_d - 1$, where we set $A^{\text{IES1}}_{d,n} = 0$ for $n < 0$. Summing $a^{\text{IES1}}_{d,m,n,\boldsymbol{s}}$ over all eligible $m \in \Phi_d$ and each n-element path $\boldsymbol{s}$ from the set $\Psi(\Gamma_m - \{d\}, n)$, we obtain

$$a^{\text{IES1}}_{d,n} = \sum_{m \in \Phi_d, n_m > n} \sum_{\boldsymbol{s} \in \Psi(\Gamma_m - \{d\}, n)} a^{\text{IES1}}_{d,m,n,\boldsymbol{s}} \tag{9}$$

for $n = 0, 1, \ldots, k^*_d - 1$, where $a^{\text{IES1}}_{d,m,0,\emptyset} = A_m/n_m$. For $n = 0, 1, \ldots, k^*_d - 1$, in each level n, $B^{\text{IES1}}_{d,n}$ is obtained by

$$B^{\text{IES1}}_{d,n} = \mathbf{E}(A^{\text{IES1}}_{d,n}, N_d). \tag{10}$$

With probability $B^{\text{IES1}}_{d,n-1} - B^{\text{IES1}}_{d,n-2}$, all servers of server group d at level $n-1$ are busy and the most senior call it is serving is a call with $|\Delta| = n - 1$. The offered traffic $A^{\text{IES1}}_{d,n-2}$ (made up of calls with $|\Delta| \le n - 2$) to the server group is overflowed with information exchange and forms the overflow traffic $v^{\text{IES1}}_{d,n}$. On the other hand, with probability $B^{\text{IES1}}_{d,n-1}$, all servers of server group d at level $n-1$ are busy, serving calls with $|\Delta| \le n - 1$. The offered traffic $a^{\text{IES1}}_{d,n-1}$ to the server group is simply overflowed without information exchange and contributes to the overflow traffic $v^{\text{IES1}}_{d,n}$. Thus, for $n = 1, 2, \ldots, k^*_d$, we derive $v^{\text{IES1}}_{d,n}$ as

$$\begin{aligned} v^{\text{IES1}}_{d,n} &= A^{\text{IES1}}_{d,n-2}(B^{\text{IES1}}_{d,n-1} - B^{\text{IES1}}_{d,n-2}) + a^{\text{IES1}}_{d,n-1} B^{\text{IES1}}_{d,n-1} \\ &= A^{\text{IES1}}_{d,n-1} B^{\text{IES1}}_{d,n-1} - A^{\text{IES1}}_{d,n-2} B^{\text{IES1}}_{d,n-2} \end{aligned} \tag{11}$$

where we set $A^{\text{IES1}}_{d,n} = 0$ and $B^{\text{IES1}}_{d,n} = 0$ for $n < 0$. Accordingly, the probability $b^{\text{IES1}}_{d,n}$ is derived by $b^{\text{IES1}}_{d,n} = v^{\text{IES1}}_{d,n+1}/a^{\text{IES1}}_{d,n}$, $n = 0, 1, \ldots, k^*_d - 1$.

The offered traffic $a^{\text{IES1}}_{s_n,m,n-1,\boldsymbol{s}-\{s_n\}}$ to server group s_n, which has overflowed sequentially from $n-1$ server groups along the path $\boldsymbol{s}$, is overflowed with probability $b^{\text{IES1}}_{s_n,n-1}$ from server group s_n and then becomes the overflow traffic $v^{\text{IES1}}_{s_n,m,n,\boldsymbol{s}}$. With probability $1/(n_m - n)$, the overflow traffic $v^{\text{IES1}}_{s_n,m,n,\boldsymbol{s}}$ will be offered to server group d, given that it is not in the path $\boldsymbol{s}$ and therefore has not been attempted. Accordingly, we have

$$a^{\text{IES1}}_{d,m,n,\boldsymbol{s}} = \frac{v^{\text{IES1}}_{s_n,m,n,\boldsymbol{s}}}{n_m - n} \tag{12}$$

and $v^{\text{IES1}}_{s_n,m,n,\boldsymbol{s}} = a^{\text{IES1}}_{s_n,m,n-1,\boldsymbol{s}-\{s_n\}} b^{\text{IES1}}_{s_n,n-1}$.

It follows that the overall blocking probability of the original model obtained by IESA1 is given by

$$B_{\text{IESA1}} = 1 - \frac{\sum_{d \in \mathcal{D}} A^{\text{IES1}}_{d,k^*_d-1}(1 - B^{\text{IES1}}_{d,k^*_d-1})}{A}. \tag{13}$$

In (13), $1 - B^{\text{IES1}}_{d,k^*_d-1}$ is the probability that server group d in IES1 has one or more servers idle at level $k^*_d - 1$. Therefore, $A^{\text{IES1}}_{d,k^*_d-1}(1 - B^{\text{IES1}}_{d,k^*_d-1})$ is the total traffic carried by server group d in IES1 estimated by IESA1. The blocking probability of calls from source m is given by

$$\hat{B}^{\text{IES1}}_m = \frac{\sum_{\boldsymbol{s} \in \Psi(\Gamma_m, n_m)} v^{\text{IES1}}_{s_{n_m},m,n_m,\boldsymbol{s}}}{A_m}. \tag{14}$$

The multi-level traffic hierarchy allows IESA1 to compute $A^{\text{IES1}}_{d,n}$ and $B^{\text{IES1}}_{d,n}$ in each level n iteratively with the initial condition $A^{\text{IES1}}_{d,0} = \sum_{m \in \Phi_d} A_m/n_m$. Thus, a unique solution for B_{IESA1} is obtained after a bounded number of iterations.

C. IESA2

As in IES1, classification of calls in IES2 according to their seniority creates a hierarchy where the j-th level of the hierarchy includes calls for which the value of Ω is j or lower. However, the seniority of a call in service can reach up to $k-1$ in any server group of IES2.

IESA2 is based on the following multi-level traffic hierarchy considered for each server group in IES2 to capture the exchange of the third attribute while retaining the first and second attributes. At level 0, the offered traffic to server group d is made up of new calls (with $|\Delta| = 0$ and $\Omega = 0$) only. An arriving call that finds the server group available is admitted; otherwise, the call is rejected. At each subsequent level $j = 1, 2, \ldots, k-1$, the offered traffic to server group d is made up of the offered traffic to it at level $j-1$ plus all calls that overflow at level $j-1$ from other server groups and hence are offered as calls with $|\Delta| = 1, 2, \ldots,$ or $\min(j, k^*_d - 1)$ and $\Omega = j$ to server group

d. An arriving call, with $|\Delta| = 0, 1, \ldots,$ or $\min(j, k_d^* - 1)$ and $\Omega = |\Delta|, |\Delta| + 1, \ldots,$ or j, that finds the server group available is admitted; otherwise, all servers of the server group are busy, serving calls with $|\Delta| = 0, 1, \ldots,$ or $\min(j, k_d^* - 1)$ and $\Omega = |\Delta|, |\Delta| + 1, \ldots,$ or j. If the seniority of the arriving call is lower than that of the most senior call in service, the arriving call overflows with information exchange; otherwise, it overflows without information exchange. Calls from source m that overflow at level j, $j = n_m - 1, n_m, \ldots,$ or $k - 1$ may become calls with $|\Delta| = n_m$ and/or $\Omega = k$. In such situations, they are by definition blocked and cleared from the system.

In each level of the analysis, IESA2 assumes that the offered traffic to each server group in IES2 is Poisson and the states of different server groups are mutually independent. Thus, the system of k server groups is decoupled into k independent subsystems, where the d-th subsystem is treated as an $M/M/N_d/N_d$ system.

For each $d \in \mathcal{D}$ in the surrogate IES2 model, define

- $a^{\text{IES2}}_{d,m,j,n,\boldsymbol{s}}$ – Offered traffic to server group d ($d \notin \boldsymbol{s}$) made up of calls from source m with $|\Delta| = n$ and $\Omega = j$ that have overflowed sequentially from n server groups in Γ_m along the path $\boldsymbol{s} = \{s_1, s_2, \ldots, s_n\}$, $m \in \Phi_d$, $j = 0, 1, \ldots, k-1$, $n = 0, 1, \ldots, \min(j, n_m - 1)$.
- $a^{\text{IES2}}_{d,j,n}$ – Offered traffic to server group d made up of calls with $|\Delta| = n$ and $\Omega = j$, $j = 0, 1, \ldots, k-1$, $n = 0, 1, \ldots, \min(j, k_d^* - 1)$.
- $\tilde{a}^{\text{IES2}}_{d,m,j,n,\boldsymbol{s}}$ – $\tilde{a}^{\text{IES2}}_{d,m,j,n,\boldsymbol{s}} = \sum_{i=n}^{j} a^{\text{IES2}}_{d,m,i,n,\boldsymbol{s}}$, $m \in \Phi_d$, $j = 0, 1, \ldots, k-1$, $n = 0, 1, \ldots, \min(j, n_m - 1)$.
- $A^{\text{IES2}}_{d,j}$ – Offered traffic to server group d at level j, $j = 0, 1, \ldots, k-1$.
- $v^{\text{IES2}}_{d,m,j,n,\boldsymbol{s}}$ – Overflow traffic from server group d ($d \in \boldsymbol{s}$) made up of calls from source m with $|\Delta| = n$ and $\Omega = j$ that have overflowed sequentially from n server groups in Γ_m along the path $\boldsymbol{s}$ ($s_n = d$), $m \in \Phi_d$, $j = 1, 2, \ldots, k$, $n = 1, 2, \ldots, \min(j, n_m)$.
- $z^{\text{IES2}}_{d,m,j,n,\boldsymbol{s}}$ – Blocked traffic from server group d ($d \in \boldsymbol{s}$) made up of calls from source m with $|\Delta| = n$ and $\Omega = j$ that have overflowed sequentially from n server groups in Γ_m along the path $\boldsymbol{s}$ ($s_n = d$), $m \in \Phi_d$, $j = 1, 2, \ldots, k$, $n = 1, 2, \ldots, \min(j, n_m)$.
- $B^{\text{IES2}}_{d,j}$ – Probability that all servers of server group d at level j are busy, serving calls with $|\Delta| = 0, 1, \ldots,$ or $\min(j, k_d^* - 1)$ and $\Omega = |\Delta|, |\Delta| + 1, \ldots,$ or j, $j = 0, 1, \ldots, k-1$.

By definition, we have $A^{\text{IES2}}_{d,j} = A^{\text{IES2}}_{d,j-1} + \sum_{n=0}^{\min(j,k_d^*-1)} a^{\text{IES2}}_{d,j,n}$, $j = 0, 1, \ldots, k-1$, where we set $A^{\text{IES2}}_{d,j} = 0$ for $j < 0$. Summing $a^{\text{IES2}}_{d,m,j,n,\boldsymbol{s}}$ over all eligible $m \in \Phi_d$ and each n-element path $\boldsymbol{s}$ from the set $\Psi(\Gamma_m - \{d\}, n)$, we obtain

$$a^{\text{IES2}}_{d,j,n} = \sum_{m \in \Phi_d, n_m > n} \sum_{\boldsymbol{s} \in \Psi(\Gamma_m - \{d\}, n)} a^{\text{IES2}}_{d,m,j,n,\boldsymbol{s}} \tag{15}$$

for $j = 0, 1, \ldots, k-1$, $n = 0, 1, \ldots, \min(j, k_d^* - 1)$, where $a^{\text{IES2}}_{d,m,0,0,\emptyset} = A_m/n_m$, $a^{\text{IES2}}_{d,m,j,0,\emptyset} = 0$ for $j = 1, 2, \ldots, k-1$.

For $j = 0, 1, \ldots, k-1$, in each level j, $B^{\text{IES2}}_{d,j}$ is obtained by

$$B^{\text{IES2}}_{d,j} = \mathbf{E}(A^{\text{IES2}}_{d,j}, N_d). \tag{16}$$

With probability $B^{\text{IES2}}_{s_n,j-1} - B^{\text{IES2}}_{s_n,j-2}$, all servers of server group s_n at level $j-1$ are busy and the most senior call it is serving is a call with $\Omega = j-1$. The offered traffic $\tilde{a}^{\text{IES2}}_{s_n,m,j-2,n-1,\boldsymbol{s}-\{s_n\}}$ (made up of calls from source m with $\Omega \le j-2$) to server group s_n, which has overflowed sequentially from $n-1$ server groups along the path $\boldsymbol{s}$, is overflowed with information exchange and forms the overflow traffic $v^{\text{IES2}}_{s_n,m,j,n,\boldsymbol{s}}$. On the other hand, with probability $B^{\text{IES2}}_{s_n,j-1}$, all servers of server group s_n at level $j-1$ are busy, serving calls with $\Omega \le j-1$. The offered traffic $a^{\text{IES2}}_{s_n,m,j-1,n-1,\boldsymbol{s}-\{s_n\}}$ to the server group is simply overflowed without information exchange and contributes to the overflow traffic $v^{\text{IES2}}_{s_n,m,j,n,\boldsymbol{s}}$. Thus, we derive $v^{\text{IES2}}_{s_n,m,j,n,\boldsymbol{s}}$ as

$$\begin{aligned} v^{\text{IES2}}_{s_n,m,j,n,\boldsymbol{s}} &= \tilde{a}^{\text{IES2}}_{s_n,m,j-2,n-1,\boldsymbol{s}-\{s_n\}}(B^{\text{IES2}}_{s_n,j-1} - B^{\text{IES2}}_{s_n,j-2}) \\ &\quad + a^{\text{IES2}}_{s_n,m,j-1,n-1,\boldsymbol{s}-\{s_n\}} B^{\text{IES2}}_{s_n,j-1} \\ &= \tilde{a}^{\text{IES2}}_{s_n,m,j-1,n-1,\boldsymbol{s}-\{s_n\}} B^{\text{IES2}}_{s_n,j-1} \\ &\quad - \tilde{a}^{\text{IES2}}_{s_n,m,j-2,n-1,\boldsymbol{s}-\{s_n\}} B^{\text{IES2}}_{s_n,j-2} \end{aligned} \tag{17}$$

for $j = 1, 2, \ldots, k$ and $n = 1, 2, \ldots, \min(j, n_m)$, where we set $\tilde{a}^{\text{IES2}}_{d,m,j,n,\boldsymbol{s}} = 0$ for $j < n$ and $B^{\text{IES2}}_{d,j} = 0$ for $j < 0$.

Recalling (1), we let $|\Delta| = n$ and $\Omega = j$ and denote $P_{m,n,j}$ for an (I, Δ, Ω)-call from source m to represent the probability P defined in (1). With probability $P_{m,n,j}$, the remaining $n_m - n$ server groups in Γ_m are all presumed to be busy, so the overflow traffic $v^{\text{IES2}}_{s_n,m,j,n,\boldsymbol{s}}$ is blocked and cleared from the system. Thus, for $j = 1, 2, \ldots, k$ and $n = 1, 2, \ldots, \min(j, n_m)$, we derive the blocked traffic $z^{\text{IES2}}_{s_n,m,j,n,\boldsymbol{s}}$ as

$$z^{\text{IES2}}_{s_n,m,j,n,\boldsymbol{s}} = v^{\text{IES2}}_{s_n,m,j,n,\boldsymbol{s}} P_{m,n,j}. \tag{18}$$

On the other hand, with probability $1 - P_{m,n,j}$, the overflow traffic $v^{\text{IES2}}_{s_n,m,j,n,\boldsymbol{s}}$ will continue to attempt a server group in $\Gamma_m - \Delta$. In particular, it is offered to server group d with probability $1/(n_m - n)$, given that it is not in the path $\boldsymbol{s}$ and therefore has not been attempted. Accordingly, we have

$$a^{\text{IES2}}_{d,m,j,n,\boldsymbol{s}} = \frac{v^{\text{IES2}}_{s_n,m,j,n,\boldsymbol{s}}}{n_m - n}(1 - P_{m,n,j}). \tag{19}$$

Similarly to (13), the overall blocking probability of the original model obtained by IESA2 is derived as

$$B_{\text{IESA2}} = 1 - \frac{\sum_{d \in \mathcal{D}} A^{\text{IES2}}_{d,k-1}(1 - B^{\text{IES2}}_{d,k-1})}{A} \tag{20}$$

where $A^{\text{IES2}}_{d,k-1}(1 - B^{\text{IES2}}_{d,k-1})$ is the total traffic carried by server group d in IES2 estimated by IESA2. The blocking probability of calls from source m is given by

$$\hat{B}^{\text{IES2}}_m = \frac{\sum_{n=1}^{n_m} \sum_{\boldsymbol{s} \in \Psi(\Gamma_m, n)} \sum_{j=n_m}^{k} z^{\text{IES2}}_{s_n,m,j,n,\boldsymbol{s}}}{A_m}. \tag{21}$$

The multi-level traffic hierarchy allows IESA2 to compute $A^{\text{IES2}}_{d,j}$ and $B^{\text{IES2}}_{d,j}$ in each level j iteratively with the initial condition $A^{\text{IES2}}_{d,0} = \sum_{m \in \Phi_d} A_m/n_m$. Thus, a unique solution for B_{IESA2} is obtained after a bounded number of iterations.

D. Scalability

In (2), (9), (15), to compute the offered traffic of a particular call type to each server group in the system, all three approximations require path enumeration of a set of permutations as a result of random hunting. In large systems, such path enumeration requiring exponentially increasing computer effort would make the algorithms computationally infeasible.

For this concern, we improve the approximations by introducing an alternative hunting scheme as a substitute of random hunting. This scheme is designed in a way that requires significantly less effort in path enumeration while retaining a certain randomness in resource selection. Under this scheme, a request from source m is allowed to attempt any server group in Γ_m with equal probability as its first choice. Then, the request follows a strictly *round-robin* (RR) order in attempting subsequent server groups in Γ_m during its hunt for an available server group. For example, given $\Gamma_m = \{d_1, d_2, \ldots, d_{n_m}\}$, if the request attempts server group d_3 as its first choice, it must follow the RR order $\{d_3, d_4, \ldots, d_{n_m}, d_1, d_2\}$ and attempt each subsequent server group with probability **one** until it finds an available one. If all server groups in the RR order have been attempted and found unavailable, the request is blocked and cleared from the system.

Define $\mathcal{R}(\Gamma_m, n)$, $n = 0, 1, \ldots, n_m$, as the set of all n-element RR paths enumerated from the set of server groups in Γ_m. Clearly, for $n > 0$, the size of the set $\mathcal{R}(\Gamma_m, n)$ is simply n_m regardless of the value of n. Define $\mathcal{R}_d(\Gamma_m - \{d\}, n)$, $n = 0, 1, \ldots, n_m - 1$, as the set of all n-element RR paths enumerated from the set of server groups in $\Gamma_m - \{d\}$ such that, after the request from source m has attempted the n server groups along an n-element RR path $\boldsymbol{s}$, it will next attempt server group d (with probability one) following the RR order of the server groups in Γ_m. Clearly, for $n > 0$, there is only one such path, so the size of the set $\mathcal{R}_d(\Gamma_m - \{d\}, n)$ is always one regardless of the value of n. By definition, $\mathcal{R}(\Gamma_m, 0) = \mathcal{R}_d(\Gamma_m - \{d\}, 0) = \emptyset$.

In forming the modified approximations, for which we use the acronym EA-RR, IESA1-RR and IESA2-RR, respectively, we replace $\Psi(\Gamma_m - \{d\}, n)$ in (2), (9), (15) with $\mathcal{R}_d(\Gamma_m - \{d\}, n)$, and $\Psi(\Gamma_m, n)$ in (8), (14), (21) with $\mathcal{R}_d(\Gamma_m, n)$. We also replace $1/(n_m - n)$ in (3), (12), (19) and $1/(n_m - i)$ in (4), (5) with unity, since there are no alternative choices in the revised hunting scheme.

VI. Numerical Results

We have conducted extensive experiments under a wide range of system parameters. Here, due to space limitation, we provide results for 1) a small system with $M = 100$, $k = 10$ and 2) a large system with $M = 500$, $k = 100$. In both systems, we set $N_d = 30$, $d \in \mathcal{D}$. In the small system, for each value of k^* ranging from 3 to 10, we set up 200 experiments; in each experiment, we set $n_m = k^*$ for $1 \le m \le 40$, $n_m = k^* - 1$ for $41 \le m \le 70$, and $n_m = k^* - 2$ for $71 \le m \le 100$. For each source m, the set Γ_m is randomly (uniformly) chosen from the set $\mathcal{D}$. The offered traffic is set at $A = 280$ in all cases. In the large system, for each value of k^* ranging from 20 to 80 at a step of 10, we set up 200 experiments; in each experiment, for each source m, we randomly choose n_m from $[k^* - 10, k^* + 10]$ and the set Γ_m from $\mathcal{D}$. The offered traffic A is chosen such that the overall blocking probability of the system is around 0.5%.

In all experiments, we obtain the blocking probability of the original model by simulation. The results are obtained in the form of an observed mean from multiple independent runs of the corresponding experiment. The confidence intervals at the 95% level based on the Student's t-distribution are maintained within $\pm 1\%$ of the observed mean. We compute the error between the approximation and the simulation in terms of the *relative error* and the *logarithmic error*. Given an approximation result x and a simulation result y, the relative error is $(x-y)/y$, and the logarithmic error is $\log_{10} x - \log_{10} y$.

For the small system, at each k^*, the mean and standard deviation of approximation errors for overall blocking probabilities of the system are obtained from the 200 experiments. The results are presented in Fig. 1(a) for the relative error. In all cases, we observe that the RR-based algorithms yield similar blocking probabilities. This demonstrates that the proposed alternative hunting scheme is an appropriate substitute of random hunting.

For the large system, at each k^*, we obtain the mean and standard deviation of approximation errors for overall blocking probabilities of the system from the 200 experiments. The results are presented in Fig. 1(b) for the relative error and in Fig. 1(c) for the logarithmic error. The approximation errors for per-source blocking probabilities are presented in the form of cumulative distribution in Fig. 1(d) for the relative error and in Fig. 1(e) for the logarithmic error. Running times of each algorithm on an Intel machine are plotted in Fig. 1(f) on a logarithmic scale.

We observe that IESA1 is very accurate when k^* is close to k. However, when k^* is small, it significantly overestimates overall blocking probabilities by up to 200%. The error of IESA1 for per-source blocking probabilities can be more than 500% as bad. EA, on the other hand, significantly underestimates the blocking probabilities as k^* increases. The almost 100% underestimation of the results means that EA predicts a result lower by orders of magnitude. This effect is demonstrated in the form of logarithmic errors in Fig. 1(c), where we see that the performance of EA is exponentially worse as k^* increases. In all cases, IESA2 is consistently better than EA, and is robust. Even in the large system, the relative error of IESA2 falls within a very small range, from -30% to -38%, for different k^*. The range of errors of IESA2 for per-source results is between -16% and -62%.

The results in Fig. 1(f) demonstrate that IESA2 is more efficient than EA with respect to the CPU running time. Given that EA is known to be scalable, so is IESA2. In fact, IESA2 is considerably more efficient than EA as k^* becomes large, corresponding to a large system, where computational efficiency of approximations becomes important. Although IESA1 is far more efficient than IESA2, it is accurate only in certain special cases, and is far from robust as observed in Fig. 1(a)–(e).

VII. Conclusion

We have introduced IESA as a new approximation methodology to address the challenging problem of blocking probability evaluation in non-hierarchical models of overflow loss

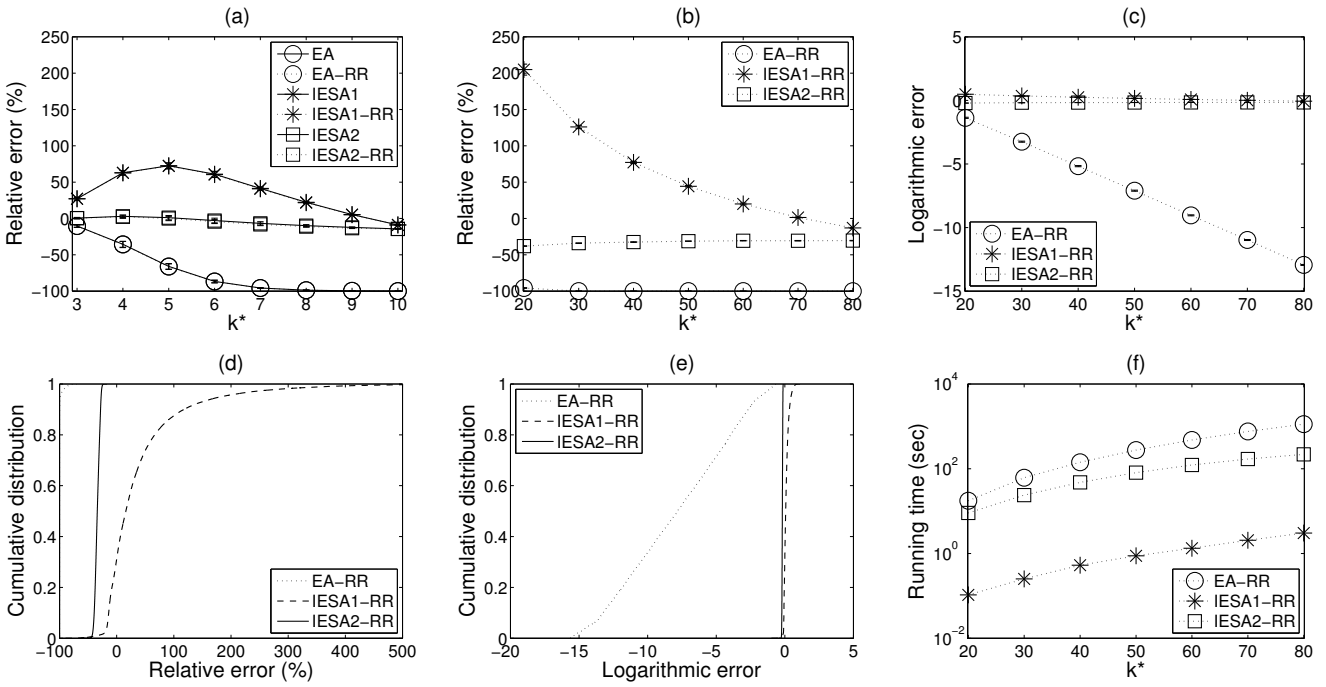

Fig. 1. Comparison of approximations. (a) Relative errors for overall blocking probabilities in the small system. (b)–(f) are results of the large system. (b) Relative errors for overall blocking probabilities. (c) Logarithmic errors for overall blocking probabilities. (d) Relative errors for per-source blocking probabilities. (e) Logarithmic errors for per-source blocking probabilities. (f) CPU running times.

systems. Extensive and statistically reliable experiments have demonstrated that the IESA2 algorithm is significantly and consistently more accurate and yet computationally more efficient, compared to the conventional EA algorithm. IESA is versatile and can be applied to more complex models of overflow loss systems. This will greatly facilitate dimensioning in telecommunication systems, resource management of video server systems, staff configuration in call centers, capacity planning for health care systems, all of which are important applications of overflow loss systems known in the literature.

REFERENCES

[1] P. K. Das and D. G. Smith, "Design of gradings subjected to unbalanced offered traffic," *IEE Proc.*, vol. 127, no. 1, pp. 41–48, Feb. 1980.

[2] P. K. Das and D. G. Smith, "Analysis of blocking probabilities in skipped gradings which are offered unbalanced traffic," *IEE Proc.*, vol. 127, no. 6, pp. 430–438, Dec. 1980.

[3] A. Girard, *Routing and Dimensioning in Circuit-Switched Networks*. Addison-Wesley, 1990.

[4] G. R. Ash, *Dynamic Routing in Telecommunications Networks*. McGraw-Hill, 1997.

[5] Z. Rosberg, H. L. Vu, M. Zukerman, and J. White, "Performance analyses of optical burst-switching networks," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 7, pp. 1187–1197, Sep. 2003.

[6] Z. Rosberg, A. Zalesky, H. L. Vu, and M. Zukerman, "Analysis of OBS networks with limited wavelength conversion," *IEEE/ACM Trans. Netw.*, vol. 14, no. 5, pp. 1118–1127, Oct. 2006.

[7] E. W. M. Wong and S. C. H. Chan, "Performance modeling of video-on-demand systems in broadband networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 7, pp. 848–859, Jul. 2001.

[8] J. Guo, E. W. M. Wong, S. Chan, P. Taylor, M. Zukerman, and K. S. Tang, "Performance analysis of resource selection schemes for a large scale video-on-demand system," *IEEE Trans. Multimedia*, vol. 10, no. 1, pp. 153–159, Jan. 2008.

[9] G. Koole and J. Talim, "Exponential approximation of multi-skill call centers architecture," in *Proc. QNETs*, 2000, pp. 23.1–23.10.

[10] P. Chevalier and N. Tabordon, "Overflow analysis and cross-trained servers," *International Journal of Production Economics*, vol. 85, no. 1, pp. 47–60, Jul. 2003.

[11] N. Litvak, M. van Rijsbergen, R. J. Boucherie, and M. van Houdenhoven, "Managing the overflow of intensive care patients," *Eur. J. Oper. Res.*, vol. 185, no. 3, pp. 998–1010, Mar. 2008.

[12] M. Asaduzzaman, T. J. Chaussalet, and N. J. Robertson, "A loss network model with overflow for capacity planning of a neonatal unit," *Annals of Operations Research*, vol. 178, no. 1, pp. 67–76, Jul. 2010.

[13] V. B. Iversen, *Teletraffic Engineering Handbook*. Tech. Univ. Denmark, 2010.

[14] R. C. McNamara, "Applications of spanning trees to continuous-time Markov processes, with emphasis on loss systems," Ph.D. dissertation, The University of Colorado, 2004.

[15] R. B. Cooper and S. S. Katz, "Analysis of alternate routing networks with account taken of the nonrandomness of overflow traffic," Bell Telephone Lab. Memo., Tech. Rep., 1964.

[16] F. P. Kelly, "Blocking probabilities in large circuit-switched networks," *Adv. Appl. Prob.*, vol. 18, no. 2, pp. 473–505, Jun. 1986.

[17] E. Brockmeyer, "A survey of A. K. Erlang's mathematical works," *Trans. Danish Acad. Tech. Sci.*, vol. 1948, no. 2, pp. 101–126, 1948.

[18] R. I. Wilkinson, "Theories of toll traffic engineering in the USA," *Bell Syst. Tech. J.*, vol. 35, no. 2, pp. 421–514, Mar. 1956.

[19] A. A. Fredericks, "Congestion in blocking systems – a simple approximation technique," *Bell Syst. Tech. J.*, vol. 59, no. 6, pp. 805–827, Jul./Aug. 1980.

[20] E. W. M. Wong, A. Zalesky, Z. Rosberg, and M. Zukerman, "A new method for approximating blocking probability in overflow loss networks," *Computer Networks*, vol. 51, no. 11, pp. 2958–2975, Aug. 2007.

[21] J. M. Holtzman, "Analysis of dependence effects in telephone trunking networks," *Bell Syst. Tech. J.*, vol. 50, no. 8, pp. 2647–2662, Oct. 1971.

[22] A. G. Hart and S. Martínez, "Sequential iteration of the Erlang fixed-point equations," *Information Processing Letters*, vol. 81, no. 6, pp. 319–325, Mar. 2002.