IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 14 May 2023, accepted 4 June 2023, date of publication 7 June 2023, date of current version 22 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3283803

SURVEY

A Century-Long Challenge in Teletraffic Theory: Blocking Probability Evaluation for Overflow Loss Systems With Mutual Overflow

ERIC W. M. WONG[1], **(Senior Member, IEEE), AND YIN-CHI CHAN**[2], **(Member, IEEE)**
[1]Department of Electrical Engineering, City University of Hong Kong, Hong Kong, SAR, China
[2]Institute for Manufacturing, University of Cambridge, CB3 0FS Cambridge, U.K.

Corresponding author: Eric W. M. Wong (eeewong@cityu.edu.hk)

This work was supported in part by the Health and Medical Research Fund of Hong Kong under Grant 16171921, and in part by the Research Grants Council (RGC) of Hong Kong under GRC General Research Fund under Grant 11104620, Grant 11102421, and Grant 11101422.

ABSTRACT In this review, we describe historical and recent developments towards tackling a century-long challenge in teletraffic theory, namely the evaluation of blocking probability in overflow systems with mutual overflow. Such systems have many applications in a variety of telecommunications and service systems, including wireless communications, cloud computing, intensive care, and emergency services, and various methods have been developed over the past century to address this challenge. In particular, the recent development of the Information Exchange Surrogate Approximation (IESA) (Wong et al., Sep. 2013; Chan and Wong, 2018) provides significantly increased accuracy and robustness compared to previous approximation methods of its kind while also providing high computational efficiency not available via simulation or exact analysis. To the best of our knowledge, IESA is the first analytical method to combine high levels of accuracy, robustness, and computational efficiency when evaluating blocking probability in overflow systems with mutual overflow, and thus forms a major breakthrough in this century-long effort.

INDEX TERMS Teletraffic theory, blocking probability, loss systems, mutual overflow.

I. INTRODUCTION

Overflow loss systems (OLSs) are an important class of stochastic models which arise in a wide variety of teletraffic and service systems applications, including wireless communications, cloud computing, intensive care, and emergency services. OLSs are defined by a set of request types (each with a given arrival process), a set of server groups (each of which serves some subset of the request types in the system), and an *overflow* policy for directing arriving requests from one server group to another until an available server is found, upon which the arriving request is assigned to that server. Alternatively, if *all* possible server groups are fully occupied at the time of the request's arrival, the arriving request is blocked and cleared from the system. The probability of a request being blocked and cleared, known as the *blocking probability*, is a key performance metric of OLSs.

Resource planning, resource allocation, and optimal resource usage in OLSs often require accurate and efficient methods for blocking probability evaluation. In many practical cases, such OLSs are not amenable to scalable blocking probability calculation as they exhibit significant state dependencies, making the state space of the system too large for exact analysis and exacerbating the problem of finding scalable and robust approximations. This problem is especially important in resource optimization, where rapid yet accurate blocking probability evaluation of a large number of candidate system configurations is key to an efficient optimization algorithm. Meanwhile, conventional estimation methods are "either time-consuming (like the discrete event simulation) or not accurate enough (like the Erlang fixed point approximation" [1] (these two methods are described in Sections II-B and II-D of this paper, respectively).

In this paper, we review various techniques for blocking probability evaluation in OLSs. The scope of this review includes the following:

The associate editor coordinating the review of this manuscript and approving it for publication was Wenchi Cheng.

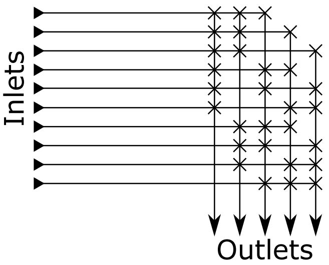


FIGURE 1. Simplified depiction of a grading from classical telephony. Note that the offered traffic to the grading does not need to be equally distributed among the outlets, creating blocking probability evaluation challenges as described by Lotze [26].

1) We describe historical developments towards tackling a century-long challenge in teletraffic theory, namely the evaluation of blocking probability in overflow systems with mutual overflow. *Mutual overflow* [2], [3], [4], [5], [6] refers to a situation where congestion in a specific server group causes overflow to the other server groups, which in turn become congested and yield overflow back to the original server group. This creates specific challenges toward the evaluation of such OLSs, as explained in Section I-B.
2) We compare a variety of methods that have been developed over the past century to address this challenge (item 1) and explain why they did not successfully tackle this challenge in terms of accuracy and computational efficiency.
3) We describe a recent method, which was developed after a century of effort on this challenge since Gray's original design [7] for a grading system. The method is called the "Information Exchange Surrogate Approximation" (IESA) [8] and provides significantly increased accuracy and robustness compared to previous approximation methods of its kind while also providing high computational efficiency not available via simulation or exact analysis. To the best of our knowledge, IESA is the first analytical method to combine high levels of accuracy, robustness, and computational efficiency when evaluating blocking probability in overflow systems with mutual overflow, and thus forms a major breakthrough in this century-long effort.
4) We explain how and why IESA works, introduce its evolution process, and describe its future development trend.

A. FROM "OLD" TELEPHONY TO WIRELESS AND BEYOND: A CENTURY OF TELETRAFFIC THEORY

The problem of accurate blocking probability evaluation in OLSs stretches back over a century. An important early example of OLSs is that of electromechanical telephone switches from the late 19th and early 20th centuries, with Gray receiving the first patent for a "grading system" in 1911 [7]. Generally speaking, a grading system [26] is a configuration of inlets and outlets in a telephone switching system where each inlet is only connected to some of the outlets. An example grading system is shown in Fig. 1.

Nearly sixty years after Gray's patent, Lotze [26] listed a number of grading-related problems that remained open, including the development of "*improved approximate methods for loss calculation, if unbalanced traffic is offered*." However, over a century since Gray's original design, and half a century since Lotze's survey paper, this problem remains unresolved, while its importance has grown due to new applications throughout the field of telecommunications and beyond. For example:

- To deliver ultra-reliable low-latency communication (URLLC), e.g. for an autonomous vehicle network, vehicles must use multiple nearby base stations to overcome frequent physical blockages [27]. An overflow policy captures the base-station preference order of vehicles in each location. Due to the low-latency requirement, the network is modeled as a loss network rather than one with delays.
- OLS models were applied to the bed management of intensive care networks in [12], [13], and [14], where some patients may be referred to any intensive care unit in a group. Other examples of OLS models in healthcare settings include [28], [29], [30], [31], and [32].
- OLS models can be used to model resource allocation in cloud services. For example, in "serverless" computing [15], servers are assigned on-the-fly to small, individual tasks.
- OLS models were used in [21], [22], [23], [24], and [25] to model the performance of emergency vehicular networks, where requests are served by the closest depot with an available vehicle.

Table 1 shows the correspondence between various concepts in the abstract OLS model and their counterparts in the above real-world applications.

B. HIERARCHICAL VERSUS NON-HIERARCHICAL OLSs

There are two classes of OLS models, as illustrated in Fig. 2: hierarchical and non-hierarchical. In hierarchical models, the server groups of the OLS are stratified into several tiers. New requests first attempt to access server groups from the lowest tier. If they are rejected from a given tier, they overflow and attempt to access server groups from a higher tier. Therefore, the traffic dependencies in hierarchical OLSs are bottom-up unidirectional: congestion in lower tiers can cause congestion in higher tiers, but not vice versa. For such systems, moment matching approaches (e.g. [33], [34], [35], [36], [37]) are available based on the assumption that the offered traffic to each tier can be treated independently, by matching the moments (i.e., mean, variance, and possibly skewness) of each layer's input traffic to that of the overflow traffic from the previous (lower) tier.

A more difficult and important challenge is the accurate and scalable evaluation of blocking probabilities in non-hierarchical OLSs (NH-OLSs) in which the overflow

TABLE 1. Correspondence between the abstract OLS model and real-world applications.

Application	Task / Service	Request	Server	Server group	Blocking Probability
Autonomous vehicle network	Vehicle-to-vehicle (V2V) communications [9]–[11]	A data packet	A 5G millimeter-wave channel	A 5G millimeter-wave base station	Probability that all replications of a packet are dropped / physically blocked
Intensive care	Bed management [12]–[14]	A critical care patient	An ICU bed	A hospital containing one or more ICUs	Probability that a new patient cannot be assigned an ICU bed
Cloud services	Serverless computing [15]	A compute request	A compute slot or virtual machine on a web server	A web server	Probability that the request cannot be assigned a server
Cloud services	Media hosting, e.g. video-on-demand [16]–[20]	A download or streaming request	A bandwidth slot on a web server	A web server	Probability that all web servers containing a replication of the requested file are at full capacity
Emergency vehicular dispatch	Fleet management [21]–[25]	A vehicle request	An emergency vehicle (e.g. ambulance, fire truck)	A vehicle depot	Probability that a request cannot be assigned a vehicle

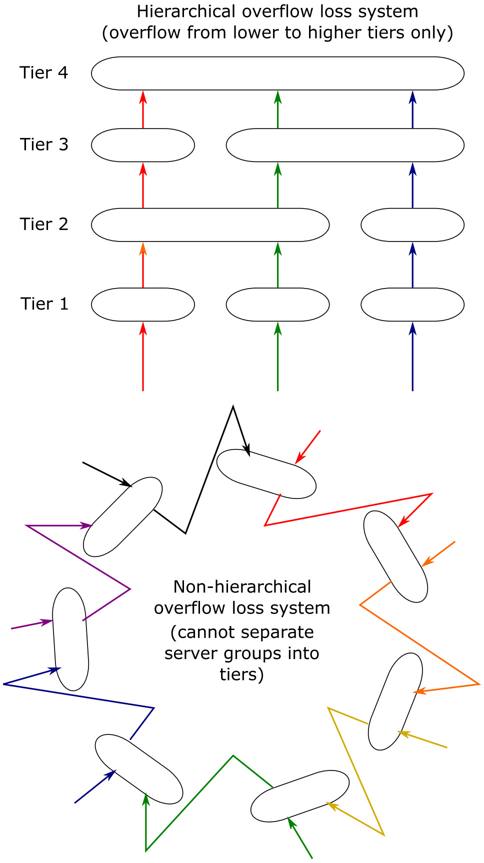


FIGURE 2. Graphical depiction of a hierarchical OLS and a non-hierarchical OLS. Each set of arrows of the same color represents an overflow sequence of server groups that requests offered to the OLS may take.

traffic from each server group may directly or indirectly affect the offered load to any other server group. In general, load dependencies in NH-OLSs are much stronger than in hierarchical OLS, and the independence assumption is more likely to lead to significant errors.

Due to this redistribution of overflow traffic throughout the OLS, resulting in more optimal resource sharing, non-hierarchical OLSs with mutual overflow in general perform *better* than hierarchical OLSs in terms of blocking probability; for example, in telephone networks, this phenomenon was associated with the transition of intercity telephony networks from a tree topology (with some additional trunk groups) to a mesh topology [38], [39], [40], [41], [42], [43]. In fact, with regard to the grading systems from which our current problem originates, Erlang [44] states that where each call may attempt up to k outlets, the *ideal* grading is one where each request attempts one out of all the possible permutations of k outlets from the set of all outlets, with equal probability (note that such an arrangement may not be possible in practice). This arrangement has the added benefit that all states with the same number of busy outlets are statistically equivalent, making Erlang's ideal grading (EIG) one of the few NH-OLSs with mutual overflow for which a scalable exact solution exists for blocking probability evaluation. The formula for the blocking probability of an EIG is known as Erlang's interconnection formula (EIF).

On the other hand, NH-OLSs in real-world applications generally deviate quite far from EIG, due to heterogeneous loading, server group availabilities, and/or server group sizes. Therefore, they do not possess simple exact solutions such as EIF for evaluation of their blocking probabilities. The difficulty in obtaining accurate and scalable approximations for such systems, as illustrated by Lotze's stated problem [26] remaining open over a century since the introduction of the grading, can be explained by the mutual traffic dependencies between server groups caused by mutual overflow. This prevents the existence of an exact product-form solution, in which the state probability distribution of a system (i.e., the full OLS) is the product of the state probability distributions of its sub-components (i.e., the server groups). Without a product-form solution, blocking probability evaluation in NH-OLSs suffers from the "curse of dimensionality": the number of possible states of an NH-OLS increases exponentially with the number of server groups in the system. Therefore, exact analysis of the state space of an NH-OLS is not a scalable method of blocking probability evaluation.

In 2013 [8], a novel methodology called the Information Exchange Surrogate Approximation (IESA) was invented, forming a major breakthrough towards an accurate, robust, and computational efficient method for blocking probability evaluation in NH-OLSs with mutual overflow. Numerical results [8], [45], [46] using IESA demonstrate increased accuracy and robustness compared to all other existing approximation methods when applied to such systems. More recently, a hybrid approach [47] was introduced combining IESA with neural networks, yielding a more accurate and robust approximation than either approach alone.

C. ORGANIZATION

The remainder of this paper is as organized as follows. In Section II, we briefly describe historical methods for blocking probability evaluation in NH-OLS, and explain why none of them achieve the full trifecta of accuracy, robustness, and computational efficiency. In Section III, we describe the development of the IESA methodology [8], [45] and its underlying principles; in particular, Section III-C, provides a numerical example to demonstrate the performance of IESA, Section III-E describes how IESA has been applied to various applications, while Section III-F describes several extensions to the original IESA algorithm (as described in [8]). Section IV highlights potential directions for the future development of IESA. Finally, some concluding remarks are given in Section V.

II. EXISTING METHODOLOGIES FOR BLOCKING PROBABILITY EVALUATION IN NH-OLSs

In this section, we describe several different existing methodologies for blocking probability evaluation in NH-OLSs. These are summarized in Table 2.

A. EXACT SOLUTION

For NH-OLSs possessing the Markov property, meaning that the next state of the system depends only on the current state (and not on previous states or elapsed time), the exact probability of each state can be evaluated by solving a system of linear equations, from which the request blocking probability can be obtained. However, as mentioned in the introduction, this is not a viable method for most NH-OLSs due to the curse of dimensionality, where the system state space is exponential with respect to the system size (i.e. number of server groups).

B. SIMULATION-BASED METHODOLOGY

There are two main branches of simulation methodology for the performance evaluation of discrete-state stochastic systems such as OLSs: discrete-event simulation (DES) [51], [52], [53] and Markov-chain simulation (MCS) [54]. In DES, a sorted list of pending events is maintained. In each iteration of the simulation loop, the DES algorithm advances the simulation clock to the time of the next pending event, removes that event from the list, and processes the event, possibly updating the system state and/or generating new events to be added to the event list. DES was a major motivator behind the development of object-oriented programming; in particular, the programming language SIMULA, released in the 1960s by the Norwegian Computing Centre, was designed specifically with DES in mind [55]. Modern software libraries for DES include SimPy, salabim [56], JaamSim [57], and simmer [58].

In contrast, in MCS, it is assumed that the system can be modeled as a Markov process, with state transition probabilities that depend only on the current state (and not on previous states or elapsed time). Simulation is conducted by random sampling of the possible next states of the system at every step, and no simulation clock is required. The simplicity of MCS means that it can easily be performed manually for small-scale systems, for example using a roulette wheel or dice.

MCS is generally more computationally efficient than DES, as it does not require a sorted event list. However, MCS can still require a significant amount of computation time. Therefore, computationally-efficient approximate analytical methods have also been developed for the evaluation of OLSs. One well-known methodology for these analytical methods is called *decomposition*, which we describe in Section II-D. Numerical results in [45] show that such methodology generally performs several orders of magnitude faster than simulation.

C. ERLANG'S INTERCONNECTION FORMULA AND EXTENSIONS

EIF works by computing the probability that, given a certain number Ω of busy servers in the system, a new arrival limited to k random attempts will encounter busy servers in all k attempts and therefore be blocked from the system. For Erlang's Ideal Grading (EIG) and a given value of Ω, this probability is equal for all arrivals:

$$P_{k,\Omega,G} = \begin{cases} \dfrac{\binom{\Omega}{k}}{\binom{G}{k}}, & k \le \Omega \le G \\ 0 & 0 \le \Omega < k, \end{cases} \tag{1}$$

where G is the total number of server groups. Defining π_n to be the probability of a state with n busy servers and A to be the offered load to the EIG (in Erlangs), we obtain the detailed balance equations [59]

$$\pi_n A \left(1 - P_{k,n,G}\right) = \pi_{n+1}(n+1) \tag{2}$$

$$\sum_{n=0}^{G} \pi_n = 1, \tag{3}$$

from which the blocking probability π_G of the EIG can be easily found. To explain (2), note that for each state n, the

TABLE 2. Summary of methods for evaluating blocking probability in NH-OLSs.

	Exact solution	Simulation	Erlang Interconnection Formula [44]	Neural networks	EFPA (e.g., [48])	IESA [8], [45]
Computational efficiency	Very low	Low	Very high	**Training**: low (requires simulation to generate training data); **Prediction**: high	High	High
Accuracy	Exact	Very high	Depends on system similarity to EIG [26]	High within range of training set, low outside range of training set [49]	Low [8], [50]	High

probability flow from state n to state $n+1$ is the probability π_n, times the offered load A, times the probability that each incoming request in state n is accepted, i.e., $\left(1-P_{k,n,G}\right)$. Conversely, the probability flow from state $n+1$ to state n is the probability π_{n+1}, times $n+1$ (i.e., the number of busy servers in state $n+1$). Detailed balance dictates that the probability flow between each pair of states in the system must be equal [59], from which (2) emerges. For additional detail on the derivation of EIF, see [60, §6.2.2].

Longley [61] found explicit alternative formulas for $P_{k,n,G}$ for certain non-ideal small systems, as well as approximations for some larger cases. Lotze [26] summarizes some other extensions to EIF, including modified terms based on a geometric series, a related approximation method for an EIG offered Engset traffic (a counterpart to Poisson traffic where the arrival rate is proportionally reduced when individuals enter the system), and the consideration of gradings with delays.

Stasiak [62] proposed an approximation method extending EIF to the case with multichannel traffic streams, where each request type may have its own service rate and seize multiple servers from the same group simultaneously. In [20] and [63], EIF-derived approximations were applied to blocking probability evaluation of video-on-demand systems; in particular, [63] considered the case where different videos have different availabilities. The technique was further extended in [64], [65] to support BPP (Binomial-Poisson-Pascal) arrival traffic. Additional applications of the technique include a cellular network model in [66] and an optical network node in [67].

On the other hand, since the methods in this subsection all assume that the probability that a new arrival is blocked depends solely on the current number of busy servers in the system, they are not accurate for systems with heterogeneous loads. As mentioned in Section I-B, NH-OLSs in real-world applications generally deviate quite far from EIG and thus their blocking probabilities cannot be estimated accurately using EIF.

D. DECOMPOSITION-BASED METHODOLOGY

To our best knowledge, decomposition-based methodology is so far the *only* scalable analytical solution for general NH-OLSs. Such methodology approximates the performance of NH-OLSs by treating each server group in an NH-OLS as an independent, full-availability queue (i.e. each request may attempt all servers in that queue). The most famous decomposition-based approach is the Erlang Fixed-Point Approximation (EFPA) [48], in which the offered traffic to each server group, including overflow traffic from other groups, is treated as if it were Poisson. The combination of the above *independence* and *Poisson* simplifying assumptions means that EFPA can be implemented simply through repeated applications of the classic Erlang B formula, and is thus quite computationally efficient.

Specifically, in EFPA, the probability of each server being fully occupied (i.e., each server is busy) is a function of the offered load to that server group; however, due to mutual overflow, circular dependencies exist between the offered loads to the server groups in the NH-OLS. Nevertheless, Brouwer's fixed-point theorem [68] can be used to show that a solution to the related blocking-probability equations always exists. These equations, and the overall blocking probability of the NH-OLS, can be evaluated using sequential fixed-point iteration [69], [70].

Variations of EFPA have been developed for sequential routing, random routing, and least-busy-first routing; for example, [71] applied EFPA to least-busy-first routing in a wired telecommunications network, while [17] applied the same in the context of a video-on-demand service. In [72], a decomposition-based approximation was derived for a special case of an OLS with delays (however, an alternative method is proposed in its place). In [19], EFPA was extended to the case of processor-sharing queues, in the context of a video-on-demand system. Multi-rate traffic, in which requests may request multiple servers in a group simultaneously, is considered in [73].

However, the fixed-point equations of EFPA can lead to multiple solutions in some cases [48], [74], [75].

Furthermore, for many types of OLSs (both hierarchical[1] [37] and non-hierarchical [8], [50]), EFPA's simplifying assumptions can cause it to underestimate blocking probability by several orders of magnitude. For hierarchical OLS, errors caused by the Poisson assumption can be reduced by taking higher moments of the overflow traffic into account (e.g. [33], [34], [35], [36], [37], [76]); however, for NH-OLSs, the independence assumption forms the dominant source of error [50]. Therefore, for NH-OLSs with Poisson arrivals, the addition of moment matching alone yields only marginal improvement over EFPA. In Section III, we describe a novel approach using the decomposition-based methodology for NH-OLSs called the *Information Exchange Surrogate Approximation* (IESA) [8], [45] that addresses the errors caused by EFPA's independence assumption. This greatly improves the accuracy and robustness of the new framework over EFPA, while still remaining computationally efficient.

An EFPA-based decomposition method was deployed in [30] for estimating rejection probabilities in a perinatal network. To improve accuracy, some subnetworks (representing multiple wards in the same hospital) were evaluated together, rather than assuming independence between all wards. Nevertheless, it was noted that some calculated probabilities "are not very close to the observed values".

Other decomposition-based approximation methods for overflow loss networks include slice methods [77], [78], [79] and a contour method [78]. These methods are based on approximating the probability density of the *number* of occupied channels on each link, rather than simply the probability that the link is fully occupied, and are generally more accurate than EFPA when applied to network applications. Finally, [80], [81] directly consider dependency effects between links in a circuit-switched network. However, these approaches are specific to networks with multi-link paths and do not apply to the generic NH-OLS model described in this paper.

1) EXTENSIONS TO EFPA FOR SYSTEMS WITH DELAYS

Overflow loss systems have been used in the literature to approximate the behavior of systems with delays, such as call centers, with some success [82]. One model for call centers is that of [83], in which an EFPA-like approximation method is proposed. An extension to this method [84] allows each call to wait for service at its last-choice server group; this is used to approximate the performance of the case where delayed calls wait in a common buffer.

2) EXTENSIONS TO EFPA FOR CELLULAR AND WIRELESS NETWORKS

Two unique properties among cellular and wireless networks that are not present in the simple overflow loss model are that of locality and handover. Locality means that a request originating from a particular cell may only obtain a channel from that cell and its immediate neighbors. Handover means that a request may move to a new cell mid-service. These requests must then be assigned a new channel in a similar manner to that of a new request.

EFPA was applied to a cellular network model with handover in [85]. In this model, handovers and call completion are modeled as competing processes, both with exponentially distributed holding times. Additionally, channel reservation is also modeled, in which the last few free channels at each base station are reserved for handover calls, thus reducing the call dropping probability (at the cost of increased blocking of fresh calls). EFPA was applied to a multi-layer cellular network in [86], with extensions to take the variance of overflow traffic into account. In this multi-layer model, there are multiple overlapping layers of differently-sized cells (possibly representing different radio frequencies and/or wireless protocols), and calls that cannot be served by a given layer may overflow to the next layer in the hierarchy. A similar model is considered in [87] where one of the layers does not form a connected graph, i.e., some group of cells are completely isolated from others.

The IESA approximation described in Section III has also been extended for cellular and wireless networks. Details are given in Sections III-E2 and III-F2. Numerical results (e.g., [46]) show that IESA for cellular networks (IESA-CN) is generally much more accurate than EFPA, especially in cellular networks with low (below 10^{-3}) rejection rates where EFPA can underestimate the true rejection rate by multiple orders of magnitude.

3) THE HYPERCUBE MODEL AND RELATED APPROXIMATIONS

The hypercube model was introduced by Larson [21], [88] to model emergency vehicle dispatches. Requests from each geographic region have a fixed preference of servers (police cars, ambulances, or otherwise), and the objective is to compute per-region and per-server statistics, e.g. the mean travel distance and the proportion of requests served by the most-favored server. The model is named for the fact that each server has two states, creating a state space that forms a $\{0, 1\}^N$ hypercube, where N denotes the number of servers.

Larson also proposed an approximation method for his hypercube model [22], which uses decomposition like EFPA, but adds correction factors to counter the effect of the independence assumption. These factors are based on considering a single, simple N-server full-availability queue and computing the ratio between the computed blocking probabilities as evaluated using the basic Erlang B formula [89] and by assuming full independence among all servers in the queue.

However, these correction factors assume a full-availability system, meaning that each request may attempt every server in the system. Furthermore, the correction factors also assume that the offered traffic is balanced among all the geographic regions, and Larson [22] noted a decrease in accuracy when this was not the case.

[1] Although fixed-point iteration is not required for hierarchical OLSs, we retain the name EFPA for consistency with NH-OLS nomenclature. Other names include "exponential decomposition" (ED), e.g. [37].

Larson's approximation was extended in [90], to handle different mean service times for each region-server pair. A further extension [25] replaces each server with a multi-server group, providing a new set of correction factors. This contrasts with an earlier method [23] which allows for preference ties between servers but continues to treat each server separately and uses the original correction factors from [22]. However, both methods still assume full accessibility.

Finally, an approximation based on the hypercube model for limited-availability systems was proposed in [24]; however, this approximation only applies to servers arranged in a ring topology, with requests limited to the nearest two servers in the ring.

4) PRODUCT-FORM UPPER BOUNDS

Van Dijk et al. [91], [92] note that OLSs with *call packing*, where in-progress requests in an overflow network are immediately rerouted to the most-preferred available server group when it becomes available, form a good surrogate for estimating the blocking probability of the equivalent OLSs without call packing. The enforcement of call packing leads to a system where the blocking probability can be expressed in product form, i.e. as the product of the blocking probability of the individual server groups. Van Dijk and Erik van der Sluis [91] proved for a simple system with two server groups, where calls offered to Group 1 may overflow to Group 2 but not vice-versa, call packing leads to an upper bound of the blocking probability when the service rate is unchanged by overflow, thus depending only on the call type. Van Dijk and Schilstra [92] state that the upper bound is also expected to hold if overflow decreases the service rate. However, they show numerically that the bound does *not* hold if overflow increases the service rate (although this case rarely appears in real-world systems). Furthermore, Van Dijk and Schilstra [92] state that the study of more complex overflow structures, for example with "parallel" (non-hierarchical) overflow, remains a "challenging point".

E. NEURAL-NETWORK-BASED METHODOLOGY

Neural networks (NN) have been proposed as a method of blocking probability evaluation in overflow loss networks, particularly optical networks [93], [94], [95], [96], [97]. In particular, [95], [96], [97] showed that single-hidden-layer feedforward networks are sufficient to provide accurate blocking probability estimation in such networks. Futhermore, the extreme learning machine (ELM) family of NN algorithms used in [95], [96], and [97] does not rely on backpropagation, unlike in conventional NN architectures, and computationally efficient methods exist for incremental addition of hidden nodes to the NN while updating the output weights of existing hidden nodes. Other examples of using NNs for blocking probability evaluation include [1], where they are used to dynamically schedule resources for an elastic optical network.

Nevertheless, there are some well-known drawbacks of NN-based approaches. Among these, the most fundamental problem is the lack of explainability of NN output (i.e., the black box problem [98]), where there is no specific method for determining or interpreting the rationale behind decisions made by an NN. Furthermore, the NN output may be very poor for input values outside the range of the training set; in other words, NNs generally have poor extrapolation capabilities. A potential approach to improve explainability and/or extrapolation capability is the use of hybrid models, where domain-specific knowledge (i.e., teletraffic theory in the current use case) is incorporated into machine learning algorithms as prior knowledge [99], [100]; this is explored further in Section III-F2.

III. INFORMATION EXCHANGE SURROGATE APPROXIMATION (IESA)

As summarized in Table 2, with the exception of IESA, all methods considered in Section II for evaluating blocking probability in NH-OLSs have drawbacks in terms of low computational efficiency or low accuracy/robustness. In particular, exact analysis is not computationally scalable, simulation requires a significant amount of computation time, EIF is not accurate for unbalanced traffic [26], neural networks are not accurate outside the range of the training set [49], and EFPA is in general not accurate for NH-OLSs [8], [50].

In particular, the high computational efficiency but low accuracy/robustness of EFPA is due to its independence assumption, which ignores state and traffic dependencies between server groups when decomposition is applied. The purpose of IESA [8], [45] is thus to develop a decomposition-based approximation like EFPA that can preserve such dependencies under decomposition. To address errors caused by EFPA's independence assumption when applied to NH-OLSs, IESA applies decomposition to a (fictitious) *surrogate* of the original system to be evaluated. The surrogate system contains special rules regarding the overflow of requests between server groups such that dependencies between the server groups are preserved under decomposition, whereas decomposition of the original system destroys these dependencies.

A. CALL ATTRIBUTES

To introduce IESA, we first define a set of three call attributes assigned to each request in the IESA surrogate model:

- an identity attribute Δ, containing general information about the request such as traffic source and arrival time;
- a history attribute, containing a list of server groups previously attempted by the request; and
- a congestion estimate Ω, containing the estimated number of busy (fully occupied) server groups in the system.

All fresh requests to the system start with an empty history attribute and a congestion estimate of zero. All requests overflowing from a fully occupied server group add that server group to their history attribute and increment their congestion

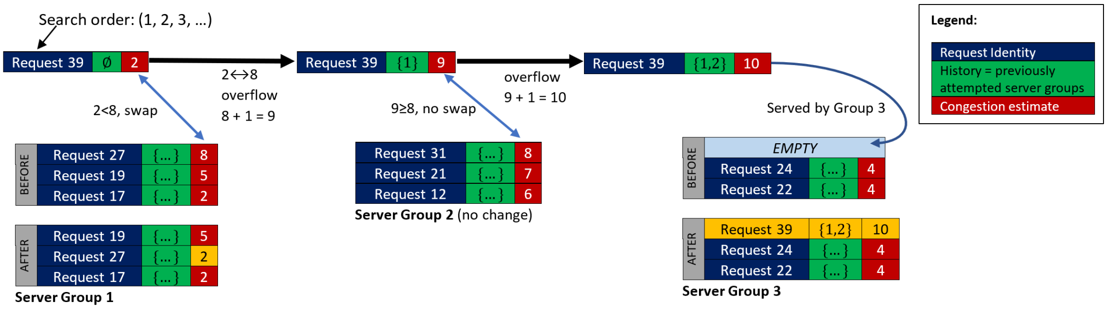


FIGURE 3. **Illustration of the information exchange mechanism in the fictitious IESA surrogate model. In this example, each server group contains three servers and can therefore handle up to three simultaneous requests.**

estimate by one. Congestion estimates can also be altered via information exchange, which we shall describe in the next subsection.

B. CONCEPTUAL DESCRIPTION AND UNDERPINNING RATIONALE

The IESA surrogate model contains an information exchange mechanism based on the following rules:

- A request arriving at a server group with at least one available server is served at that group and no information exchange occurs.
- For a request arriving at a fully occupied server group, find the most "senior" request in service at that server group, defined as the request with the highest congestion estimate (ties broken arbitrarily). The incoming request exchanges congestion estimates with this senior request *if and only if* it has a lower congestion estimate than the senior request.

An example of the information exchange mechanism in operation is shown in Fig. 3. Note that the information exchange mechanism is only used by the (fictitious) surrogate system to estimate the blocking probability for IESA and is not actually applied to the real system.

In addition to an information exchange mechanism, the IESA surrogate model has an *early abandonment* mechanism, in which a request may leave the system as a blocked request without attempting all server groups available to it. The probability of early abandonment increases with respect to the length of the request's history attribute and the value of its congestion estimate, and also depends on the total number of server groups available to the request and the total number of server groups in the NH-OLS. For the case where all requests may attempt the same number of server groups, two equations have been proposed [8], [13]:

$$P_{n,k,\Omega,G} = \begin{cases} \dfrac{\binom{\Omega-n}{k-n}}{\binom{G-n}{k-n}}, & n \leq k \leq G \\ 0 & 0 \leq n < k, \end{cases} \tag{4}$$

$$P^{+}_{n,k,\Omega,G} = \begin{cases} \dfrac{\binom{\Omega}{k-n}}{\binom{G}{k-n}}, & n \leq k \leq G \\ 0 & 0 \leq n < k, \end{cases} \tag{5}$$

where n is the number of attempted server groups so far, i.e. $|\Delta|$, k is the maximum number of server groups each request is allowed to attempt, Ω is the congestion estimate, and G is the total number of server groups in the system. Due to the nature of the information exchange mechanism, we have $\Omega \geq k$.

Note the similarities of (4) and (5) to (1). This is because assuming that the probability of selecting $k-n$ fully occupied server groups depends only on Ω and G naturally yields an EIG-like scenario where heterogeneities in the NH-OLS are ignored. (However, unlike EIF, these heterogeneities are captured elsewhere in the IESA algorithm.) Additionally, (5) results in higher blocking probability estimates than (4) except when $k = G$.

The early abandonment mechanism is thus designed such that requests most likely to abandon the system early are also the requests most likely to be blocked anyway in the original NH-OLS, where no early abandonment mechanism is applied, such that the blocking probability of the IESA surrogate model is close to (but generally slightly higher than) that of the original NH-OLS.

Note that while the information exchange and early abandonment mechanisms of IESA were designed to address errors caused by its independence assumption, they also address errors caused by the Poisson assumption. This is because the probability of early abandonment increases with respect to the congestion estimate and the number of previously attempted server groups. Therefore, the left-over non-removed traffic has a higher portion of fresh traffic, which is Poisson and independent, and hence becomes closer to Poisson and independent overall. This leads to less error when decomposition is applied to the IESA surrogate model (relative to direct decomposition of the original "true" model in EFPA), as the decomposition methodology assumes Poisson and independent input to each server group. This gives an intuition of why IESA generally produces a good estimate of blocking probability for the original NH-OLS under evaluation; this is depicted in Fig. 4. Thus the success of IESA lies in the ability to remove "unwanted" (non-Poisson and highly dependent) traffic flows from the system in such a way that (1) the total blocked and removed traffic in the surrogate system is similar to the volume of blocked traffic in the original

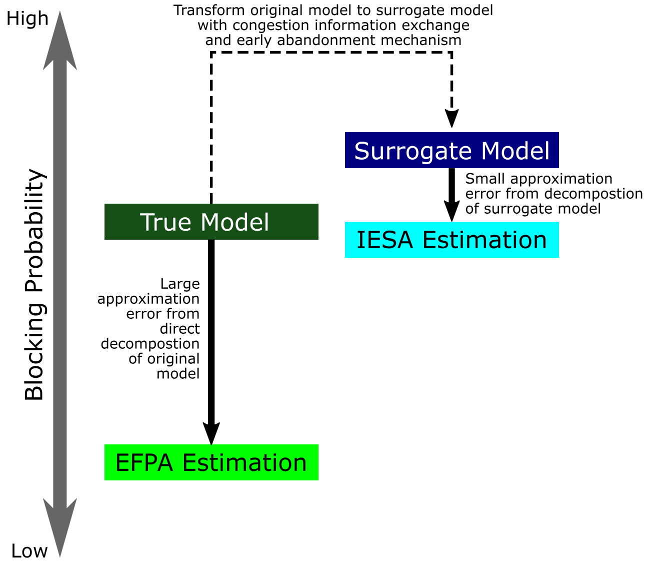


FIGURE 4. Conceptual illustration of blocking probability approximation in NH-OLSs based on the IESA framework (adapted from [45], [50]).

system, and (2) the non-removed traffic is easier to evaluate than in the original system. A more detailed explanation of this ability is given in [50, §3.1–3.2] for a predecessor to IESA, which works under the same two principles.

To further highlight how the information exchange mechanism in IESA reduces the complexity of the surrogate system, making it easier to evaluate, note that the surrogate system can be described using a hierarchical traffic model, based on the congestion estimate. This is because, due to the information exchange and early abandonment mechanisms, we are able to transform the original non-hierarchical traffic structure (containing mutual traffic dependencies) to a hierarchical traffic structure (containing one-way traffic dependencies only) with a finite number of levels, where each level corresponds to one congestion estimate level. The elegance of such a structure is that, due to one-way traffic dependence, the volume of requests to each server group at each congestion estimate level is equal to that as if all requests with a higher congestion estimate were barred from the system [45]. As a result, in IESA, we can calculate the traffic from the bottommost level first and then move up one level at one time until we reach the highest level, as in a traditional hierarchical OLS. Therefore, we can finish the calculation in a fixed, finite number of iterations, unlike EFPA, for which fixed-point iteration is generally required and the rate of convergence is not guaranteed. More specifically, since the levels of the IESA hierarchy are linked to the congestion estimate, and a higher congestion estimate is linked to both higher system congestion and traffic dependencies, the IESA surrogate model preserves this congestion and dependency information under decomposition (as it is encoded into the traffic hierarchy itself), while this information is completely destroyed under direct decomposition of the original model (as in EFPA).

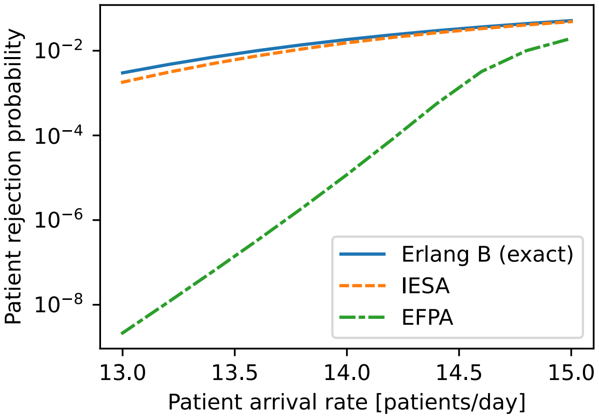


FIGURE 5. Patient rejection probability for a system of intensive care units with respect to the total patient arrival rate.

C. NUMERICAL EXAMPLE – INTENSIVE CARE

To demonstrate the impact of EFPA's approximation error in a practical setting and the improvement of IESA over EFPA, consider a system of fifteen intensive care units (ICUs) with fifteen beds each. Patients may be referred to any ICU in the system, forming a *fully-connected* NH-OLS. The rejection probability of patients due to lack of capacity can therefore be computed exactly using the Erlang B formula. Fig. 5 shows the rejection probability of the ICU system with respect to the total arrival rate of patients to the system, assuming a mean length of stay of fifteen days, as evaluated using the Erlang B formula, EFPA, and IESA. Note that EFPA shows negligible (less than 10^{-6}) patient rejection even as the actual rejection probability approaches one percent, whereas IESA provides a much more accurate estimate of the actual rejection probability, with up to six orders of magnitude of improvement over EFPA. Furthermore, a bed allocation level considered by EFPA to meet a quality of service with a one-percent rejection probability of patients will in fact see roughly four times as many patient rejections (with potentially fatal consequences), while IESA can basically meet the target quality of service.

D. THEORETICAL RESULTS

A number of theoretical results regarding IESA are given in [45] for an NH-OLS model using random routing (meaning that requests attempt server groups available to them in random order). Provided that the arrival process of fresh requests to each server group is an independent Poisson process:

- IESA (and its extensions in Section III-F) is asymptotically exact as the number of server groups increases to infinity, while the number of server groups available to each request remains fixed. Alternatively, the arrival process of fresh requests to each server group can be an independent *non-Poisson* process, as long as the node model representing each decoupled server group exactly models the arrival process of fresh requests to each server group and the service time distribution of the requests is modeled exactly. Note that this result is very general and holds even when the offered load of fresh requests to each server group, the number of servers in each group, and/or the number of server groups each request is allowed to attempt is heterogeneous. As far as we know, this is the first result on asymptotic exactness for NH-OLSs (note that Kelly's asymptotic exactness result for EFPA [48] is for a network with fixed routing, i.e., without overflow).

- For the case that the arrival process of fresh requests to each server group is an independent Poisson process, all server groups are accessible to all requests, and each server group has only one server, the IESA blocking probability estimate is always between the true blocking probability and the EFPA estimate, i.e. IESA is always at least as good as EFPA in terms of accuracy. Furthermore, under critical loading (mean arrival rate equal to the server capacity), the approximation error of IESA is bounded (in terms of the ratio between the true blocking probability and the IESA estimate) while the approximation error of EFPA is unbounded. These results were proved by showing equivalence of IESA in this special case to an earlier surrogate-based approximation [50], for which these theorems were proved in [101].

In particular, the first result above demonstrates that the asymptotic behavior of our IESA surrogate model is consistent with the asymptotic behavior of the original NH-OLS in this regime. These results thus provide some further level of justification for the design of our IESA surrogate model.

E. PRACTICAL APPLICATIONS

As shown in Table 1, the IESA can evaluate the performance of many real word applications, which can be modeled as NH-OLSs. Here we provide three examples.

1) IESA FOR AN INTENSIVE CARE NETWORK MODEL

In [13], IESA was applied to a cluster of intensive care units (ICUs). Due to the nature of intensive care patients, only certain types of patients may be referred to more than one ICU; therefore, the network contains a mix of both overflow-capable and local-only traffic. It was found that while IESA provides a good estimate of the rejection rate of overflow-capable patients, EFPA is more accurate when estimating the rejection rate of local-only patients.

However, the ICU cluster model in [13] is limited to patients with either single-ICU or all-ICU availability. A challenge for the future will be to apply IESA to systems with more complicated mixes of server availability. For example, in content delivery networks (another application of NH-OLSs) the number of replications of some content is closely related to the demand for that content. As demand for different content can vary greatly, so too can the number of file replications, i.e. the server availability for each type of request.

2) IESA FOR CELLULAR NETWORKS (IESA-CN)

In [46] and [102], IESA was applied to a cellular network model, in which requests offered to each cell may only be served by the target cell (i.e., by that cell's associated base station) or one of its neighboring cells. Due to this *locality* effect, the original IESA is not accurate for the cellular network model. A modified version of IESA, namely IESA-CN, improves the accuracy of IESA for cellular networks by modifying the probability for which early abandonment occurs after each request overflow from a fully occupied cell. Curve-fitting was employed to optimize IESA-CN for a wide range of cellular network configurations using only a single parameter. Finally, IESA-CN was applied to cellular networks with both irregular cell boundaries and loads and shown to be accurate even in this case.

3) OPTICAL NETWORKS

An earlier version of IESA, called the Overflow Priority Classification Approximation (OPCA) [50], has been successfully applied to optical networks [103], [104], [105], [106]. This includes bufferless optical networks with deflection routing, in which optical bursts/packets are permitted to overflow to an alternate trunk when all channels comprising the first-choice trunk are busy [103]. The blocking probability of such networks depends on the network size, trunk size, the maximum number of allowable deflections, and burst/packet length. In [104], theoretical bounds on the accuracy of OPCA were presented for optical burst-switching networks with deflection routing. It was also shown numerically that high accuracy can be obtained using a small number of OPCA iterations. In [105], OPCA was applied to circuit-switched optical networks with both long-lived and short-lived transmissions, where long-lived transmissions have preemptive priority over the short-lived ones. Finally, in [106], OPCA is extended to apply to multi-service multi-rate optical networks in which certain transmissions may occupy multiple bandwidth channels simultaneously on the same path. For the purpose of network dimensioning, a hybrid method using the *maximum* of EFPA and this new ''service-based OPCA'' was proposed.

F. FURTHER EXTENSIONS

In this subsection, we describe two extensions to IESA.

1) IESA WITH TWO-STREAM NODE MODEL (IESA-2S) FOR NON-POISSON INPUT TRAFFIC [45]

It has been found that in NH-OLSs with random routing, overflow traffic quickly converges to Poisson as the total number of server groups grows [45]. However, consider a modified system in which the *first* server group attempted by each request is *not* random. This aligns with real-world applications where requests will often have a preferred server group, for example one based on physical proximity. In such cases, it is important to capture the possible non-Poisson nature of fresh request arrivals. This is implemented in IESA-2S [45], which models fresh and overflow traffic to each server group ("node") separately. IESA-2S was found in [45] to be more accurate than plain IESA for NH-OLSs with both bursty and smooth fresh traffic. Note that smooth fresh traffic is another of Lotze's open challenges from [26].

Furthermore, if the node model of each server group captures the nature of the fresh requests to the NH-OLS exactly, then IESA-2S is asymptotically exact as the total number of server groups in the system, while the number of server groups available to each request remains fixed [45]. This extends a previous theoretical result [45] for IESA and EFPA which only applies to systems with Poisson fresh traffic.

Finally, as a step towards addressing Lotze's original problem regarding blocking probability evaluation in heterogeneous NH-OLSs, [45] also considers certain types of heterogeneities, including heterogeneous loading, server group availabilities, and server group sizes. In all three cases, IESA-2S was shown to be the most accurate decomposition-based approximation among all considered by a significant margin. Furthermore, the aforementioned asymptotic exactness result for IESA-2S applies even in these heterogeneous cases.

2) IESA WITH NEURAL NETWORKS (IESA+NN)

As stated previously, the success of IESA lies in whether the total blocked and removed traffic in the surrogate system is similar to the volume of blocked traffic in the original system. IESA+NN [47] further improves the accuracy and robustness of IESA by modifying the probability for which early abandonment occurs after each request overflow, as in IESA-CN, even though there is no locality in the original NH-OLS model. Additionally, while the tuning parameter in IESA-CN was fitted using classical polynomial regression, in IESA+NN, neural networks are used instead. A conceptual diagram illustrating the connection between the NN and the IESA algorithm in IESA+NN is shown in Fig. 6. Such hybrid models, where domain-specific knowledge (i.e., teletraffic theory in the current use case) is incorporated into machine learning algorithms as prior knowledge, have been previously studied in various applications throughout science and engineering [99], [100]. In such hybrid models, the neural network compensates for inaccuracies in the prior model, while the the prior model can control extrapolation in input regions lacking training data, making the model more robust. Additionally, the use of a prior model means that hybrid models can generally be trained using less data than conventional NN models [99].

The result is a highly accurate, robust, and computationally efficient (given a pre-trained neural network) algorithm, including for NH-OLSs with non-Poisson fresh traffic and non-exponential service times (as are assumed by the original version of IESA). Moreover, IESA+NN was found to be more robust than direct neural-network-based approximation of NH-OLS blocking probability when applied to parameter ranges outside of that covered by the training set, due to the bounds set by the IESA portion of the approximation, which is guided by the underlying teletraffic theory. Note that since the neural network in IESA+NN tunes a parameter that is *specific* to IESA's early abandonment mechanism, there is no equivalent EFPA+NN.

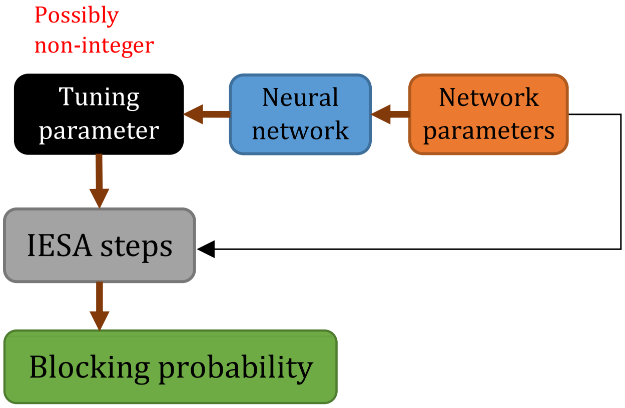


FIGURE 6. Conceptual digram for IESA+NN.

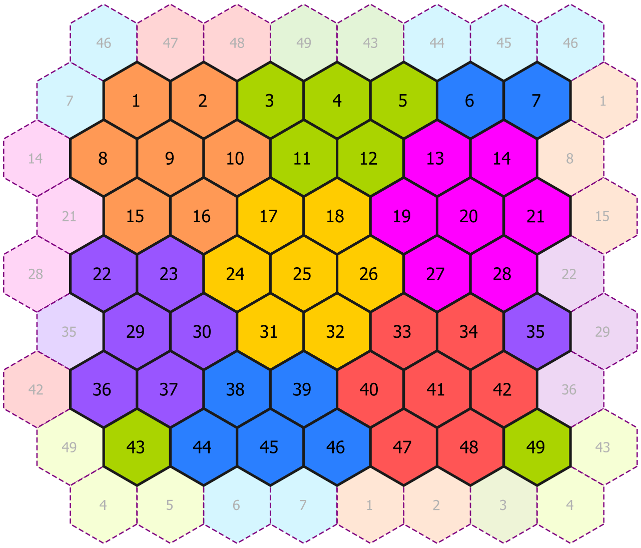
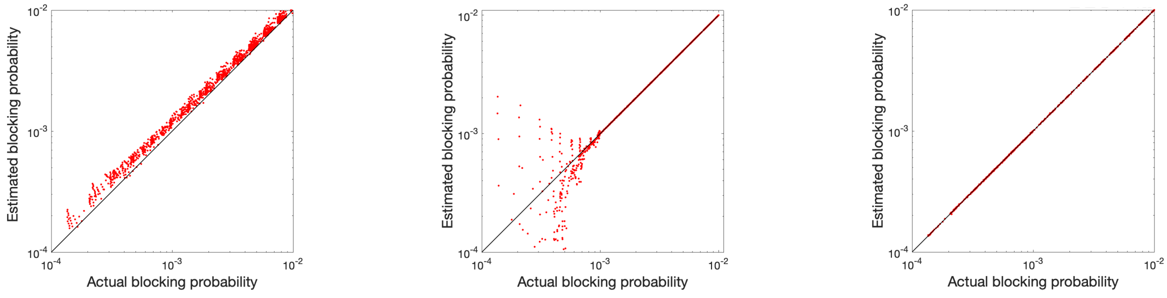


FIGURE 7. 49-cell wraparound (toroidal) network topology. Reproduced from [49].

IESA+NN has also been applied to cellular and wireless networks and shown to be more accurate and robust than both IESA-CN and the direct neural-network-based approach [49]. To demonstrate the application of IESA+NN to cellular networks, we consider a 49-cell wraparound (toroidal) network, as shown in Fig. 7. All cells support up to 10 simultaneous requests and receive the same traffic load except for a central cluster of seven cells, each of which receives α times the regular load ($\alpha = 0.8$ to 2).

We compare IESA-CN [46], IESA+NN for cellular networks [49], and direct NN evaluation of the blocking probability, with a blocking probability range of 10^{-2} to 10^{-3} for the training set and 10^{-2} to 10^{-4} for the training set. The results, shown in Fig. 8, demonstrate that IESA-CN is moderately accurate across the entire range of the training set, whereas direct NN fails to extrapolate to blocking



(a) IESA-CN (b) Direct NN (c) IESA-NN

FIGURE 8. Blocking probability results for a cellular network model. Subfigure (b) shows the poor extrapolation ability of the direct NN approach. Reproduced from [49].

probabilities outside the training set. Finally, IESA+NN is more accurate and robust than the other two approximations, producing reliable blocking probability estimates across the entire training-set range.

IV. DIRECTIONS FOR THE FUTURE DEVELOPMENT OF IESA

In this section we describe potential directions for the future development of IESA:

A. APPLYING IESA TO OTHER APPLICATIONS

IESA can be applied to other important applications in which the system can be modeled as an NH-OLS. However, each new application has its own challenges for applying IESA to. For example, consider an ICU network in a metropolitan city, composed of multiple connected clusters. This creates a two-layer patient flow architecture with both intra-cluster (local) and inter-cluster transfer of patients based on ICU occupancy, in contrast with the single-cluster model previously considered in [13] (see Section III-E1). Note that IESA has already been applied previously to multi-layer NH-OLSs with both "vertical" and "horizontal" overflow [107], but without multiple patient types. Furthermore, the approximation errors of IESA in [107], while much less than for EFPA, are still quite large in many cases, highlighting the need for further improvement.

B. IMPROVING THE ACCURACY OF IESA AND/OR IESA+NN

We can further improve the accuracy of IESA by, for example, developing new surrogate systems to replace the current system. This involves a trade-off between improving the accuracy of IESA and maintaining computational efficiency. For example, we may replace the (scalar) congestion estimate Ω in the current surrogate system with the *set* of server groups in the system believed to be busy (either via direct observation or information exchange). Preliminary results by the authors of this review suggest that such modification improves the accuracy of IESA slightly in NH-OLSs with heterogeneous loads, at the cost of increased computational complexity.

Another approach is to examine the effect of different neural network algorithms when used in IESA+NN, improving upon existing IESA+NN implementations as described in Section III-F2. For example, one limitation of current IESA+NN implementations [47], [49] is that the inputs to the NN are based on system-wide average values; thus heterogeneities in the network (e.g., in terms of the offered load or number of neighbors of each cell) are not captured by the NN part of IESA+NN.

C. IESA AS A CONGESTION CONTROL METHOD

The information exchange mechanism in IESA's fictitious surrogate model can potentially be used as a congestion control mechanism. In particular, networks and systems where overflow traffic consumes more resources than non-overflow traffic are prone to bistability, in which the network/system gets stuck in a high-blocking state where most requests are overflow requests [74], [75]. Note that while the surrogate model in IESA, as used currently, is a fictitious model, the objective of IESA-based congestion control is to use IESA's mechanisms in the *actual* system to be controlled.

Preliminary results [108] suggest that information-exchange-based congestion control can be used on its own or in conjunction with traditional trunk reservation [3] to stabilize such networks and systems. Further investigation requires answering the following research questions:

- What indicators should be used to estimate the congestion level of the network? For example, the current IESA surrogate model assigns an Ω attribute to each request, denoting the estimated number of busy server groups in the system. Alternatively, each request may carry a *set* of server groups believed to be busy. The challenge here is to strike a balance between the effectiveness of the control mechanism and the required communication overhead in the system to be controlled.
- How to estimate the actual congestion level of the network from the values of the indicators? For example, we may assume a linear relationship.
- How to respond to the congestion? For example, we may discard new arrivals with some probability when the congestion indicator reaches a certain level (as in the current IESA surrogate model).

V. CONCLUDING REMARKS

In this paper, we described and compared different methods for the evaluation of blocking probability in NH-OLSs, an important century-long challenge in teletraffic theory with applications in various service and telecommunications systems, including wireless communications, cloud computing, intensive care and emergency services. Among these, IESA and its extensions provide a good combination of accuracy, robustness, and computational efficiency across a wide range of system parameters, thus forming a major breakthrough in this century-long effort. The introduction of machine learning in IESA+NN, as detailed in Section III-F2, provides further improved accuracy and robustness, and was shown to be more robust than a direct machine-learning-only approach. Note that since the only difference between IESA+NN and base IESA is the modification of the early abandonment mechanism, IESA+NN can be applied to any NH-OLS application where IESA and EFPA can be used, making IESA+NN a generic and versatile approach for blocking probability evaluation in NH-OLSs.

Although discrete-event simulation remains the golden standard for accurate blocking probability evaluation in stochastic systems, the availability of fast yet accurate/robust approximation frameworks such as IESA will be very useful in applications that require rapid performance evaluation of a large number of configurations, such as optimization and dynamic control. Such approximation frameworks can also be useful for obtaining a better understanding of the mechanics behind congestion in stochastic systems by reducing complex systems into their most abstract form, while simulation can be used for validation of modeling results at the expense of requiring more computation time and data.

ACKNOWLEDGMENT

The authors would like to thank the associate editor and the anonymous referees for their constructive comments and suggestions, which have helped them to improve the quality of this manuscript.

REFERENCES

[1] X. Lin, S. Lin, Y. Li, J. Shao, K. Shi, and Y. Li, "Double-machine-learning-based resource scheduling method for offloading transfers," in *Proc. 27th Asia Pacific Conf. Commun. (APCC)*, Oct. 2022, pp. 114–119.

[2] F. Le Gall and J. Bernussou, "An analytical formulation for grade of service determination in telephone networks," *IEEE Trans. Commun.*, vol. COM-31, no. 3, pp. 420–424, Mar. 1983.

[3] L. Mason, "On the stability of circuit-switched networks with non-hierarchical routing," in *Proc. 25th IEEE Conf. Decis. Control*, Dec. 1986, pp. 1345–1347.

[4] H. Inamori, "Performance evaluation of mutal overflow routing for hierarchical packet-switching networks," *Electron. Commun. Jpn.*, vol. 71, no. 6, pp. 111–122, Jun. 1988.

[5] U. R. Krieger, "Analysis of a loss system with mutual overflow in a Markovian environment," in *Numerical Solution of Markov Chains*, W. J. Stewart, Ed. Boca Raton, FL, USA: CRC Press, 1991, ch. 16, pp. 303–328, doi: 10.1201/9781003210160-16.

[6] M. Glabowski and P. Walkowiak, "Simulation studies of communication systems with mutual overflows and threshold mechanisms," in *Proc. Int. Conf. Broadband Commun. Next Gener. Netw. Multimedia Appl. (CoBCom)*, Jul. 2018, pp. 1–8.

[7] E. A. Gray, "Method and means for connecting telephone apparatus," U.S. Patent 1 002 388, Sep. 5, 1911.

[8] E. W. M. Wong, J. Guo, B. Moran, and M. Zukerman, "Information exchange surrogates for approximation of blocking probabilities in overflow loss systems," in *Proc. 25th Int. Teletraffic Congr. (ITC)*, Sep. 2013, pp. 1–9.

[9] A. Iyer, A. Kherani, A. Rao, and A. Karnik, "Secure V2V communications: Performance impact of computational overheads," in *Proc. IEEE INFOCOM Conf. Comput. Commun. Workshops*, Apr. 2008, pp. 1–6.

[10] D. M. Mughal, J. S. Kim, H. Lee, and M. Y. Chung, "Performance analysis of V2V communications: A novel scheduling assignment and data transmission scheme," *IEEE Trans. Veh. Technol.*, vol. 68, no. 7, pp. 7045–7056, Jul. 2019.

[11] A. U. Rahman, Y. Gupta, and G. Ghatak, "On the characterization of V2V link performance in highway vehicular networks," in *Proc. IEEE 3rd 5G World Forum (5GWF)*, Sep. 2020, pp. 442–447.

[12] 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.

[13] Y.-C. Chan, E. W. M. Wong, G. Joynt, P. Lai, and M. Zukerman, "Overflow models for the admission of intensive care patients," *Health Care Manag. Sci.*, vol. 21, no. 4, pp. 554–572, Dec. 2018.

[14] A. R. Rutherford, S. L. Zimmerman, M. Moeini, R. Barket, S. Ahkioon, and D. E. G. Griesdale, "Simulation model of a multi-hospital critical care network," in *Proc. Winter Simul. Conf. (WSC)*, Dec. 2022, pp. 951–960.

[15] R. Patil, T. S. Chaudhery, M. A. Qureshi, V. Sawant, and H. Dalvi, "Serverless computing and the emergence of function-as-a-service," in *Proc. Int. Conf. Recent Trends Electron., Inf., Commun. Technol. (RTEICT)*, Aug. 2021, pp. 764–769.

[16] 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.

[17] 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.

[18] J. Guo, Y. Wang, K.-S. Tang, S. Chan, E. W. M. Wong, P. Taylor, and M. Zukerman, "Evolutionary optimization of file assignment for a large-scale video-on-demand system," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 6, pp. 836–850, Jun. 2008.

[19] J. P. Munoz-Gea, S. Traverso, and E. Leonardi, "Modeling and evaluation of multisource streaming strategies in P2P VoD systems," *IEEE Trans. Consum. Electron.*, vol. 58, no. 4, pp. 1202–1210, Nov. 2012.

[20] S. Hanczewski and M. Stasiak, "Modeling of video on demand systems," in *Computer Networks* (Communications in Computer and Information Science), vol. 431. Berlin, Germany: Springer, 2014, pp. 233–242.

[21] R. C. Larson, "A hypercube queuing model for facility location and redistricting in urban emergency services," *Comput. Oper. Res.*, vol. 1, no. 1, pp. 67–95, Mar. 1974.

[22] R. C. Larson, "Approximating the performance of urban emergency service systems," *Oper. Res.*, vol. 23, no. 5, pp. 845–868, Oct. 1975.

[23] T. H. Burwell, J. P. Jarvis, and M. A. McKnew, "Modeling co-located servers and dispatch ties in the hypercube model," *Comput. Oper. Res.*, vol. 20, no. 2, pp. 113–119, Feb. 1993.

[24] J. B. Atkinson, I. N. Kovalenko, N. Kuznetsov, and K. V. Mykhalevych, "A hypercube queueing loss model with customer-dependent service rates," *Eur. J. Oper. Res.*, vol. 191, no. 1, pp. 223–239, Nov. 2008.

[25] S. Budge, A. Ingolfsson, and E. Erkut, "Technical note—Approximating vehicle dispatch probabilities for emergency service systems with location-specific service times and multiple units per location," *Oper. Res.*, vol. 57, no. 1, pp. 251–255, Feb. 2009.

[26] A. Lotze, "History and development of grading theory," in *Proc. ITC*, 1967, pp. 148–161.

[27] J. Wu, M. Wang, Y.-C. Chan, E. W. M. Wong, and T. Kim, "Performance evaluation of 5G mmWave networks with physical-layer and capacity-limited blocking," in *Proc. IEEE 21st Int. Conf. High Perform. Switching Routing (HPSR)*, May 2020, pp. 1–6.

[28] E. L. Blair and C. E. Lawrence, "A queueing network approach to health care planning with an application to burn care in new York state," *Socio-Economic Planning Sci.*, vol. 15, no. 5, pp. 207–216, Jan. 1981.

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

[30] M. Asaduzzaman and T. J. Chaussalet, "An overflow loss network model for capacity planning of a perinatal network," *J. Roy. Stat. Soc. Ser. A, Statist. Soc.*, vol. 174, no. 2, pp. 403–417, Apr. 2011.

[31] N. Izady and I. Mohamed, "A clustered overflow configuration of inpatient beds in hospitals," *Manuf. Service Operations Manag.*, vol. 23, no. 1, pp. 139–154, Jan. 2021.

[32] X. Gong, X. Wang, L. Zhou, and N. Geng, "Managing hospital inpatient beds under clustered overflow configuration," *Comput. Oper. Res.*, vol. 148, Dec. 2022, Art. no. 106021.

[33] A. Kuczura, "The interrupted Poisson process as an overflow process," *Bell Syst. Tech. J.*, vol. 52, no. 3, pp. 437–448, Mar. 1973.

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

[35] P. Chevalier and N. Tabordon, "Overflow analysis and cross-trained servers," *Int. J. Prod. Econ.*, vol. 85, no. 1, pp. 47–60, Jul. 2003.

[36] A. Brandt and M. Brandt, "Individual overflow and freed carried traffics for a link with trunk reservation," *Telecommun. Syst.*, vol. 29, no. 4, pp. 283–308, Aug. 2005.

[37] G. J. Franx, G. Koole, and A. Pot, "Approximating multi-skill blocking systems by hyperexponential decomposition," *Perform. Eval.*, vol. 63, no. 8, pp. 799–824, Aug. 2006.

[38] G. R. Ash, A. H. Kafker, and K. R. Krishnan, "Intercity dynamic routing architecture and feasibility," in *Proc. 10th Int. Teletraffic Congr.*, Montreal, QC, Canada, Jun. 1983, p. 8, Paper 2. [Online]. Available: https://gitlab2.informatik.uni-wuerzburg.de/itc-conference/itc-publications-public/-/raw/master/itc10/ash831.pdf

[39] E. W. M. Wong and T. S. Yum, "Maximum free circuit routing in circuit-switched networks," in *Proc. IEEE INFOCOM*, Jun. 1990, pp. 934–937.

[40] G. R. Ash and B. D. Huang, "An analytical model for adaptive routing networks," *IEEE Trans. Commun.*, vol. 41, no. 11, pp. 1748–1759, Nov. 1993.

[41] E. W. M. Wong, A. K. M. Chan, and T.-S.-P. Yum, "A taxonomy of rerouting in circuit-switched networks," *IEEE Commun. Mag.*, vol. 37, no. 11, pp. 116–122, Nov. 1999.

[42] E. W. M. Wong, A. K. M. Chan, and T.-S.-P. Yum, "Analysis of rerouting in circuit-switched networks," *IEEE/ACM Trans. Netw.*, vol. 8, no. 3, pp. 419–427, Jun. 2000.

[43] G. R. Ash and P. Chemouil, "20 years of dynamic routing in circuit-switched networks: Looking backward to the future," *IEEE Global Commun. Newslett.*, pp. 1–4, Oct. 2004. [Online]. Available: https://www.researchgate.net/publication/240639778

[44] A. K. Erlang, "The application of the theory of probabilities in telephone administration," in *The Life and Works of A.K. Erlang* (Transactions of the Danish Academy of Technical Sciences), no. 3, E. Brockmeyer, H. L. Halstrøm, and A. Jensen, Eds. Copenhagen, Denmark: The Copenhagen Telephone Company, 1948, pp. 201–215.

[45] Y.-C. Chan and E. W. M. Wong, "Blocking probability evaluation for non-hierarchical overflow loss systems," *IEEE Trans. Commun.*, vol. 66, no. 5, pp. 2022–2036, May 2018.

[46] J. Wu, E. W. M. Wong, J. Guo, and M. Zukerman, "Performance analysis of green cellular networks with selective base-station sleeping," *Perform. Eval.*, vol. 111, pp. 17–36, May 2017.

[47] Y.-C. Chan, E. W. M. Wong, and C. S. Leung, "Evaluating non-hierarchical overflow loss systems using teletraffic theory and neural networks," *IEEE Commun. Lett.*, vol. 25, no. 5, pp. 1486–1490, May 2021.

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

[49] Y.-C. Chan, J. Wu, E. W. M. Wong, and C. S. Leung, "Integrating teletraffic theory with neural networks for quality-of-service evaluation in mobile networks," *SSRN*, May 2023, doi: 10.2139/ssrn.4457238.

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

[51] J. Banks and J. S. Carson, "Introduction to discrete-event simulation," in *Proc. 18th Conf. Winter Simul.*, 1986, pp. 17–23.

[52] S. Robinson, "Discrete-event simulation: From the pioneers to the present, what next?" *J. Oper. Res. Soc.*, vol. 56, no. 6, pp. 619–629, Jun. 2005.

[53] A. J. Collins, F. S. A. Pour, and C. A. Jordan, "Past challenges and the future of discrete event simulation," *J. Defense Model. Simul.*, vol. 20, no. 3, pp. 351–369, Jul. 2023, doi: 10.1177/15485129211067175.

[54] M. H. Ackroyd, "Increasing the efficiency of roulette simulation of teletraffic," *Electron. Lett.*, vol. 14, no. 9, p. 270, 1978.

[55] O.-J. Dahl and K. Nygaard, "SIMULA: An ALGOL-based simulation language," *Commun. ACM*, vol. 9, no. 9, pp. 671–678, Sep. 1966.

[56] R. van der Ham, "Salabim: Discrete event simulation and animation in Python," *J. Open Source Softw.*, vol. 3, no. 27, p. 767, Jul. 2018.

[57] D. H. King and H. S. Harrison, "Open-source simulation software 'JaamSim,'" in *Proc. Winter Simulations Conf. (WSC)*, Dec. 2013, pp. 2163–2171.

[58] I. Ucar, B. Smeets, and A. Azcorra, "Simmer: Discrete-event simulation for R," *J. Stat. Softw.*, vol. 90, no. 2, pp. 1–30, 2019, doi: 10.18637/jss.v090.i02.

[59] H. Chen and D. D. Yao, "Birth-death queues," in *Stochastic Modelling and Applied Probability*. New York, NY, USA: Springer, 2001, ch. 1, pp. 1–13.

[60] V. B. Iversen. (2015). *Teletraffic Engineering and Network Planning*. [Online]. Available: http://orbit.dtu.dk/files/118473571/Teletraffic_34342_V_B_Iversen_2015.pdf

[61] H. A. Longley, "The efficiency of gradings," *Post Office Electr. Eng. J.*, vol. 41, pp. 45–49, Jan. 1948.

[62] M. Stasiak, "An approximate model of a switching network carrying mixture of different multichannel traffic streams," *IEEE Trans. Commun.*, vol. 41, no. 6, pp. 836–840, Jun. 1993.

[63] S. Hanczewski and M. Stasiak, "Performance modelling of video-on-demand systems," in *Proc. 17th Asia Pacific Conf. Commun.*, Oct. 2011, pp. 784–788.

[64] M. Glabowski, M. Sobieraj, and M. Stasiak, "Modelling limited-availability groups with BPP traffic and bandwidth reservation," in *Proc. 5th Adv. Int. Conf. Telecommun.*, 2009, pp. 89–94.

[65] M. Glabowski, S. Hanczewski, M. Stasiak, and J. Weissenberg, "Modeling Erlang's ideal grading with multirate BPP traffic," *Math. Problems Eng.*, vol. 2012, pp. 1–35, Jan. 2012.

[66] M. Glabowski, S. Hanczewski, and M. Stasiak, "Modelling of cellular networks with traffic overflow," *Math. Problems Eng.*, vol. 2015, pp. 1–15, May 2015.

[67] S. Hanczewski, M. Stasiak, and J. Weissenberg, "Non-full-available queueing model of an EON node," *Opt. Switching Netw.*, vol. 33, pp. 131–142, Jul. 2019.

[68] L. E. J. Brouwer, "Uber abbildung von mannigfaltigkeiten," *Mathematische Annalen*, vol. 71, no. 1, pp. 97–115, Mar. 1911.

[69] F. E. Browder and W. V. Petryshyn, "The solution by iteration of nonlinear functional equations in Banach spaces," *Bull. Amer. Math. Soc.*, vol. 72, no. 3, pp. 571–575, 1966.

[70] A. Hart, "Sequential iteration of the Erlang fixed-point equations," *Inf. Process. Lett.*, vol. 81, no. 6, pp. 319–325, Mar. 2002.

[71] S.-P. Chung, A. Kashper, and K. W. Ross, "Computing approximate blocking probabilities for large loss networks with state-dependent routing," *IEEE/ACM Trans. Netw.*, vol. 1, no. 1, pp. 105–115, Feb. 1993.

[72] S. Chan, "Least loaded sharing in fog computing cluster," in *Proc. 15th Int. Conf. Netw. Services*, 2019, pp. 27–31.

[73] S.-P. Chung and K. W. Ross, "Reduced load approximations for multirate loss networks," *IEEE Trans. Commun.*, vol. 41, no. 8, pp. 1222–1231, Aug. 1993.

[74] R. J. Gibbens, P. J. Hunt, and F. P. Kelly, "Bistability in communication networks," in *Disorder in Physical Systems*, G. Grimmett and D. Welsh, Eds. Oxford, U.K.: Oxford Univ. Press, 1990, pp. 113–128.

[75] D. Martirosyan and P. Robert, "The equilibrium states of large networks of Erlang queues," *Adv. Appl. Probab.*, vol. 52, no. 2, pp. 617–654, Jun. 2020.

[76] R. B. Cooper and S. S. Katz, "Analysis of alternate routing networks with account taken of nonrandomness of overflow traffic," Bell Telephone Lab., Murray Hill, NJ, USA, Memorandum MM64-3122-2, Nov. 1964. [Online]. Available: https://www.cse.fau.edu/~bob/publications/Cooper_&_Katz_1964.pdf

[77] K. Jung, Y. Lu, D. Shah, M. Sharma, and M. S. Squillante, "Revisiting stochastic loss networks: Structures and algorithms," in *Proc. ACM SIGMETRICS Int. Conf. Meas. Model. Comput. Syst.*, Jun. 2008, pp. 407–418.

[78] J. Anselmi, Y. Lu, M. Sharma, and S. Squillante, "Improved approximation for the Erlang loss model," *Queueing Syst.*, vol. 63, pp. 217–239, Dec. 2009.

[79] K. Jung, Y. Lu, D. Shah, M. Sharma, and M. S. Squillante, "Revisiting stochastic loss networks: Structures and approximations," *Math. Oper. Res.*, vol. 44, no. 3, pp. 890–918, Aug. 2019.

[80] M. Bebbington, P. Pollett, and I. Ziedins, "Two-link approximation schemes for loss networks with linear structure and trunk reservation," *Telecommun. Syst.*, vol. 19, no. 2, pp. 187–207, 2002.

[81] M. R. Thompson and P. K. Pollett, "A reduced load approximation accounting for link interactions in a loss network," *J. Appl. Math. Decis. Sci.*, vol. 7, no. 4, pp. 229–248, Jan. 2003.

[82] P. Chevalier and J.-C. Van Den Schrieck, "Optimizing the staffing and routing of small-size hierarchical call centers," *Prod. Oper. Manag.*, vol. 17, no. 3, pp. 306–319, May 2008.

[83] G. Koole and J. Talim, "Exponential approximation of multi-skill call centers architecture," in *Proc. QNETs*, Jul. 2000, pp. 1–10.

[84] A. N. Avramidis, W. Chan, and P. L'Ecuyer, "Staffing multi-skill call centers via search methods and a performance approximation," *IIE Trans.*, vol. 41, no. 6, pp. 483–497, Apr. 2009.

[85] D. McMillan, "Traffic modelling and analysis for cellular mobile networks," in *Proc. ITC*, 1991, pp. 627–632.

[86] P. Fitzpatrick, C. S. Lee, and B. Warfield, "Teletraffic performance of mobile radio networks with hierarchical cells and overflow," *IEEE J. Sel. Areas Commun.*, vol. 15, no. 8, pp. 1549–1557, Oct. 1997.

[87] G. Song, J. Wu, J. Schormans, and L. Cuthbert, "Erlang's fixed-point approximation for performance analysis of HetNets," *J. Appl. Math.*, vol. 2012, pp. 1–13, Jan. 2012.

[88] R. C. Larson, "Public sector operations research: A personal journey," *Oper. Res.*, vol. 50, no. 1, pp. 135–145, Feb. 2002.

[89] A. K. Erlang, "Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges," in *The Life and Works of A.K. Erlang* (Transactions of the Danish Academy of Technical Sciences), no. 2, E. Brockmeyer, H. L. Halstrøm, and A. Jensen, Eds. Copenhagen, Denmark: The Copenhagen Telephone Company, 1948, pp. 138–155.

[90] J. P. Jarvis, "Approximating the equilibrium behavior of multi-server loss systems," *Manag. Sci.*, vol. 31, no. 2, pp. 235–239, Feb. 1985.

[91] N. M. Van Dijk and E. Van Der Sluis, "Call packing bound for overflow loss systems," *Perform. Eval.*, vol. 66, no. 1, pp. 1–20, Jan. 2009.

[92] N. van Dijk and B. Schilstra, "On two product form modifications for finite overflow systems," *Ann. Oper. Res.*, vol. 310, no. 2, pp. 519–549, Mar. 2021.

[93] D. R. B. De Araujo, C. J. A. Bastos-Filho, and J. F. Martins-Filho, "Methodology to obtain a fast and accurate estimator for blocking probability of optical networks," *J. Opt. Commun. Netw.*, vol. 7, no. 5, pp. 380–391, May 2015.

[94] D. R. B. Araujo, C. J. A. Bastos-Filho, and J. F. Martins-Filho, "Artificial neural networks to estimate blocking probability of transparent optical networks: A robustness study for different networks," in *Proc. 17th Int. Conf. Transparent Opt. Netw. (ICTON)*, Jul. 2015, pp. 1–4.

[95] H. C. Leung, C. S. Leung, E. W. M. Wong, and S. Li, "Extreme learning machine for estimating blocking probability of bufferless OBS/OPS networks," *J. Opt. Commun. Netw.*, vol. 9, no. 8, pp. 682–692, Aug. 2017.

[96] S. Li, H. C. Leung, E. W. M. Wong, and C. S. Leung, "Enhancement of extreme learning machine for estimating blocking probability of OCS networks with fixed-alternate routing," *IEEE Access*, vol. 7, pp. 52319–52330, 2019.

[97] S. Chakraborty, A. K. Turuk, and B. Sahoo, "MELM-GRBFNN: A modified extreme learning machine trained Gaussian radial basis function neural network model for estimating blocking probability of OBS network," in *Proc. IEEE Region Conf. (TENCON)*, Nov. 2020, pp. 478–483.

[98] W. J. von Eschenbach, "Transparency and the black box problem: Why we do not trust AI," *Philosophy Technol.*, vol. 34, no. 4, pp. 1607–1622, Sep. 2021.

[99] M. L. Thompson and M. A. Kramer, "Modeling chemical processes using prior knowledge and neural networks," *AIChE J.*, vol. 40, no. 8, pp. 1328–1340, Aug. 1994.

[100] J. Willard, X. Jia, S. Xu, M. Steinbach, and V. Kumar, "Integrating scientific knowledge with machine learning for engineering and environmental systems," *ACM Comput. Surv.*, vol. 55, no. 4, pp. 1–37, Nov. 2022.

[101] E. W. M. Wong, B. Moran, A. Zalesky, Z. Rosberg, and M. Zukerman, "On the accuracy of the OPC approximation for a symmetric overflow loss model," *Stochastic Models*, vol. 29, no. 2, pp. 149–189, Apr. 2013.

[102] J. Wu, E. W. M. Wong, Y.-C. Chan, and M. Zukerman, "Power consumption and GoS tradeoff in cellular mobile networks with base station sleeping and related performance studies," *IEEE Trans. Green Commun. Netw.*, vol. 4, no. 4, pp. 1024–1036, Dec. 2020.

[103] E. W. M. Wong, J. Baliga, M. Zukerman, A. Zalesky, and G. Raskutti, "A new method for blocking probability evaluation in OBS/OPS networks with deflection routing," *J. Lightw. Technol.*, vol. 27, no. 23, pp. 5335–5347, Dec. 2009.

[104] S. Li, M. Wang, E. W. M. Wong, V. Abramov, and M. Zukerman, "Bounds of the overflow priority classification for blocking probability approximation in OBS networks," *J. Opt. Commun. Netw.*, vol. 5, no. 4, pp. 378–393, Apr. 2013.

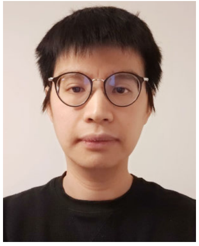
[105] M. Wang, S. Li, E. W. M. Wong, and M. Zukerman, "Blocking probability analysis of circuit-switched networks with long-lived and short-lived connections," *J. Opt. Commun. Netw.*, vol. 5, no. 6, pp. 621–640, Jun. 2013.

[106] M. Wang, S. Li, E. W. M. Wong, and M. Zukerman, "Performance analysis of circuit switched multi-service multi-rate networks with alternative routing," *J. Lightw. Technol.*, vol. 32, no. 2, pp. 179–200, Jan. 15, 2014.

[107] Y.-C. Chan, J. Guo, E. W. M. Wong, and M. Zukerman, "Surrogate models for performance evaluation of multi-skill multi-layer overflow loss systems," *Perform. Eval.*, vol. 104, pp. 1–22, Oct. 2016.

[108] E. W. M. Wong and Y.-C. Chan, "Improved performance and stability in overflow loss systems via exchange of congestion information," *TechRxiv*, Jan. 2023, doi: 10.36227/techrxiv.21908415.v1.

ERIC W. M. WONG (Senior Member, IEEE) received the B.Sc. and M.Phil. degrees in electronic engineering from The Chinese University of Hong Kong, Hong Kong, in 1988 and 1990, respectively, and the Ph.D. degree in electrical and computer engineering from the University of Massachusetts, Amherst, MA, USA, in 1994. He is currently an Associate Professor with the Department of Electrical Engineering, City University of Hong Kong, Hong Kong. His research interests include analysis and design of telecommunications and computer networks, energy-efficient data center design, green cellular networks, 5G millimeter wave communication networks, and optical networking.



YIN-CHI CHAN (Member, IEEE) received the B.Math. degree from the University of Waterloo, in 2010, and the M.Sc. and Ph.D. degrees from the City University of Hong Kong, in 2011 and 2017, respectively. He is currently a Research Associate in discrete event simulation with the Institute for Manufacturing, University of Cambridge. His current research interests include process modeling, simulation, and the analysis of healthcare systems.