Blocking Probability Evaluation for Non-Hierarchical Overflow Loss Systems

Yin-Chi Chan, *Member, IEEE*, and Eric W. M. Wong, *Senior Member, IEEE*

***Abstract*— Non-hierarchical overflow loss systems (NH-OLSs) with mutual overflow have applications in many telecommunications and service systems, e.g., cellular networks, video-on-demand, emergency healthcare, or cloud services. In this paper, we make fundamental contributions in teletraffic modeling for blocking probability evaluation in NH-OLSs. These are based on the development of the well-known decomposition methodology, which decomposes an NH-OLS into independent subsystems and applies a suitable node model to each subsystem. In particular, we provide new both theoretical and numerical results involving both new and existing approximation methods. We show that these results cover a broad range of NH-OLSs, including those with heterogeneous arrival processes, server group size, and/or routing, with both Poisson and non-Poisson arrivals of fresh requests. Our theoretical results include the first scalable asymptotic exactness results for NH-OLSs. Our new approximations include the first computationally efficient and fairly accurate approximation for NH-OLSs with both mutual overflow and non-Poisson input with asymptotic exactness properties.**

***Index Terms*— Overflow loss systems, blocking probability, mutual overflow routing, analytical approximation.**

I. Introduction

OVERFLOW loss systems (OLSs) constitute an important class of stochastic models which arise in a wide variety of telecommunications and service systems applications. OLSs are defined by a set of request types, a set of servers organized into groups, and a policy for assigning each arriving request to a server whenever possible. Requests are said to *overflow* from one server group to another until an available server is found, or are blocked and cleared from the system if all possible server groups are fully occupied at the time of the request's arrival. The probability of a request being blocked and cleared, known as the blocking probability, is an important performance metric of OLSs.

In particular, in a non-hierarchical OLS (NH-OLS), requests may attempt server groups in any arbitrary order. NH-OLSs can be applied to a large number of real world applications; for a list of examples, see Table I. However, accurate yet computationally efficient evaluation of blocking probability in NH-OLSs remains a long-standing open problem, having first been considered over a century ago in electromechanical telephone switching systems [17]–[21], but still relevant today as evidenced by the list of modern applications in Table I. The main difficultly is the "curse of dimensionality": the number of possible states of an NH-OLS increases exponentially with the number of server groups in the system. Furthermore, it is known that systems with overflow do not possess a product form solution [22], which means that exact analysis of the state space of an NH-OLS is not a scalable method of blocking probability evaluation. In particular, NH-OLSs can exhibit what is known as *mutual overflow* [23]–[25]: overload at one server group causes overflow to other server groups, in turn overloading those server groups and causing overflow back to the original server group. This causes a mutual dependency between the states of different server groups.

Additional challenges occur when the arrival process of fresh requests to the server groups is non-Poisson. This means that the time to the next arrival of a fresh request to a server group may be state-dependent. Modeling the state of each arrival process on top of that of the server groups further increases the dimensionality of the overall system state compared to a system with Poisson arrivals of fresh requests. As mutual overflow and non-Poisson arrivals are frequently observed in NH-OLSs (see examples in Sections II-A and II-B, respectively), the development of such an approximation method is of high importance.

Due to the lack of a scalable exact method for blocking probability evaluation, computer simulation is often used to evaluate the performance of various NH-OLSs. However, simulations are time-consuming, especially for large systems. This disadvantage is especially important in optimization problems where fast blocking probability evaluation of many system configurations is key to an efficient optimization algorithm. Therefore, the challenge is to construct a computationally efficient yet accurate approximation method for a wide range of NH-OLSs, including those with mutual overflow and non-Poisson input. Due to the aforementioned "curse of dimensionality", a scalable approximation of the blocking probability requires a dramatic reduction in the number of states considered at any given time.

A. Decomposition as a Methodology for Blocking Probability Evaluation in NH-OLSs

One well-known methodology for estimating blocking probability in NH-OLSs is to *decompose* the system into a set of statistically independent nodes (each representing a single

Manuscript received July 18, 2017; revised October 25, 2017; accepted December 12, 2017. Date of publication December 18, 2017; date of current version May 15, 2018. The associate editor coordinating the review of this paper and approving it for publication was D. Wu. *(Corresponding author: Yin-Chi Chan.)*

The authors are with the Department of Electronic Engineering, City University of Hong Kong, Hong Kong (e-mail: ycchan26@cityu.edu.hk; eeewong@cityu.edu.hk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCOMM.2017.2784450

TABLE I
CORRESPONDENCE BETWEEN THE ABSTRACT NH-OLS MODEL AND REAL-WORLD APPLICATIONS

Application	Request	Server	Server group	Blocking Probability	Examples
Cellular network	A voice call (traditional cellular network), web session, or data packet	A radio channel	A base station	Probability that a call/session cannot be completed, or a packet is dropped	[1]–[4]
Video-on-demand network	A download or streaming request	A download or streaming slot on a file server	A file server	Probability that a download or streaming request cannot be served	[5]–[7]
ICU network	A critical care patient	An ICU bed	An ICU ward	Probability that a new patient cannot be assigned an ICU bed	[8]
Emergency vehicular dispatch	A vehicle request	An emergency vehicle (e.g. ambulance, fire truck)	A vehicle depot	Probability that a request for a vehicle cannot be served	[9], [10]
Cloud computing	A request for a virtual machine (VM)	A VM slot	A physical machine	Probability that a new VM cannot be allocated to a physical machine	[11]–[14]
Call center	A phone call	A call center agent	A group of agents with the same skill set	Probability that a caller cannot be assigned an agent	[15], [16]

server group), applying appropriate *node models* to simplify the analysis of the offered traffic of fresh and overflow requests to each server group [15], [26], [27]. Due to the decomposition, the computational complexity of the problem can be significantly reduced. A well-known decomposition-based approximation method in the literature is the Erlang fixed-point approximation (EFPA) [15], [27], which applies decomposition to the "true" system model, treating each server group in the system as an independent Erlang B queue [28]. While EFPA is an accurate approximation for certain types of networks [27], [29], [30], it has also been shown to produce large approximation errors in NH-OLSs with mutual overflow [31]–[33].

EFPA makes three major simplifying assumptions:

1) All traffic offered to a server group, both fresh and overflow, is Poisson. In reality, overflow traffic generally has a higher peakedness (variance-to-mean ratio) than the offered traffic.
2) All traffic streams offered to a server group have the same blocking probability. In reality, peakier traffic will experience more blocking than smoother traffic [34]. Note that Assumption 1 implies Assumption 2 but not vice versa.
3) The states (number of busy servers) of the server groups are mutually independent.

Note that the first two assumptions stem from the use of the Erlang B node model, whereas Assumption 3 stems from decomposing the "true" system into independent subsystems.

Another decomposition-based method is the Information Exchange Surrogate Approximation (IESA) [4], [8], [32], [35], [36]. Unlike EFPA, which applies decomposition directly to a "true" model of the NH-OLS, IESA addresses Assumption 3 of EFPA by applying decomposition to a *surrogate* system model, which is designed to preserve state dependency information between server groups when decomposition is applied. Therefore, the approximation error caused by decomposition is much reduced compared to direct decomposition

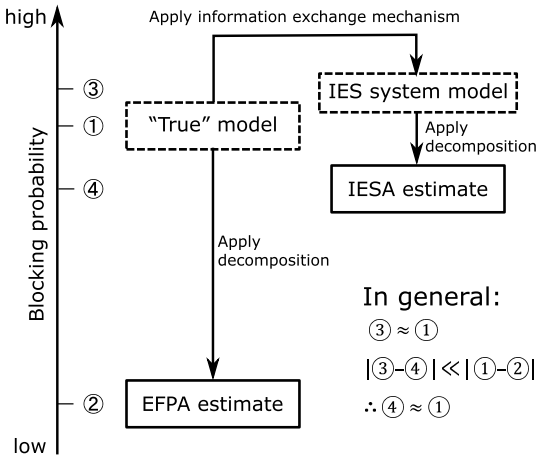


Fig. 1. Design principles of the IES system model.

of the "true" model, as depicted in Fig. 1 (see Section III-B for a detailed explanation of the IES surrogate model). However, as we shall show in this paper, both EFPA and IESA are inadequate in approximating the blocking probability of NH-OLSs offered non-Poisson fresh traffic, due to the assumptions associated with the Erlang B node model (i.e. Assumptions 1 and 2).

B. Contributions of This Paper

In this paper, we make fundamental contributions in teletraffic modeling through the development of the decomposition methodology for NH-OLSs, with new both theoretical and numerical results involving both new and existing approximation methods. We show that these results cover a broad range of NH-OLSs, including those with heterogeneous arrival processes, server group size and/or routing, with both Poisson and non-Poisson arrivals of fresh requests.

We consider a generic NH-OLS model with G server groups and random routing, where each group consists of N servers and each request may attempt a maximum k groups.

We show that the offered load to each server group composed of *overflow* requests tends towards Poisson and independent of that to any other server group as $G \to \infty$ with k fixed, as long as the arrival processes of fresh requests to each server group are mutually independent. Therefore, if the arrival processes of fresh requests to each group are also Poisson and mutually independent, then all three assumptions of EFPA, as listed in Section I-A, become true and EFPA is asymptotically exact as $G \to \infty$ with k fixed. Even in the case where fresh requests do *not* follow a Poisson process, the above result implies that *any* decomposition-based approximation is asymptotically exact for *any* arbitrary arrival process of fresh requests to the system, as long as it uses the true system model (or a model equivalent to the true model in the limiting case) and an exact node model exists for each server group in the limiting case where the overflow traffic to each server group becomes independent Poisson processes.

Using the above result, we present a limiting regime (Limiting Regime 1) where both EFPA and IESA are asymptotically exact as $G \to \infty$ with k fixed, provided that the arrival process of fresh requests to each server group is an independent Poisson process. 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, these are the first results on asymptotic exactness for NH-OLSs.

In addition, we also consider another limiting regime [33] (Limiting Regime 2) where $G = k$ and $N = 1$. We show for Poisson arrivals of fresh requests that the IESA estimate is as least as accurate as the EFPA estimate and that under critical loading, the ratio of the exact blocking probability to the IESA estimate is bounded above by $\sqrt{2}$, whereas the corresponding ratio for EFPA was shown in [33] to be unbounded.

We also extend our results to the case where the arrival process of fresh requests is non-Poisson by removing the two assumptions of both EFPA and IESA associated with the Erlang B node model. Note that existing approximations that consider both non-Poisson traffic and mutual overflow, for example [37], do not scale to a large number of overflow streams offered to the same server group. In contrast, we use a simple two-stream node model where all overflow traffic to a server group is modeled as a single Poisson process, making the node model scalable to NH-OLSs with a large number of server groups. We show that modeling overflow traffic as Poisson is asymptotically correct for NH-OLSs with random routing as $G \to \infty$ with k fixed, if the arrival processes of fresh requests to each server group is mutually independent.

By applying EFPA and IESA with the new two-stream node model, we create the new approximation methods EFPA-2S and IESA-2S, respectively. Since the new two-stream node model addresses Assumptions 1 and 2 above (those associated with the Erlang B node model) and the IES system model addresses Assumption 3, IESA-2S, which combines both the two-stream node model and the surrogate system model, depends on none of the three main simplifying assumptions associated with EFPA. In addition, the scalability of the two-stream node model means that EFPA-2S and IESA-2S are both computationally efficient. Numerical results show that IESA-2S is quite accurate even when the total number of server groups is limited, especially compared to EFPA, IESA, and EFPA-2S. As far as we know, IESA-2S is the first fairly accurate yet computationally efficient approximation for NH-OLSs with non-Poisson input. Note that EFPA-2S and IESA-2S revert back to EFPA and IESA, respectively, when the arrival process of fresh requests to each server group is Poisson, thus maintaining backwards compatibility.

We prove that EFPA-2S and IESA-2S are both asymptotically exact for Limiting Regime 1 (i.e. $G \to \infty$ with k fixed and random routing) even when the arrival process of fresh requests to each server group is non-Poisson. Similarly to our result for NH-OLSs with Poisson input, we show that this result applies even for heterogeneous systems. Furthermore, this result does not depend on the nature of the arrival process of fresh requests and holds true as long as the node model associated with the input traffic is available and exact in the limiting case, where the overflow traffic to each server group becomes independent and Poisson. This is also demonstrated numerically in this paper using interrupted Poisson processes (IPP) [38] and Engset processes [39] for the arrival processes of fresh requests to each server group.

In summary, our contributions in this paper are:

- Proof that under random routing, the arrival processes of overflow requests to each server group tend toward independent Poisson processes as $G \to \infty$ with k fixed, provided that the arrival process of fresh requests to each server groups is independent of that to the other groups.
- For NH-OLSs with Poisson arrivals of fresh requests, provided the arrival process of fresh requests to each server group is independent of that to the other groups:
  - Proof that EFPA and IESA are asymptotically exact for Limiting Regime 1, i.e. all three simplifying assumptions listed in Section I-A become true (first scalable asymptotic exactness results for NH-OLSs).
  - Proof for Limiting Regime 2 that the IESA estimate is as least as accurate as the EFPA estimate and that under critical loading, the ratio of the exact blocking probability to the IESA estimate is bounded above by $\sqrt{2}$.
- For NH-OLSs with non-Poisson arrivals of fresh requests, provided the arrival process of fresh requests to each server group is independent of that to the other groups:
  - The development of a scalable two-stream node model for use with EFPA and IESA (thus producing new computationally-efficient approximations, namely EFPA-2S and IESA-2S).
  - Proof that EFPA-2S and IESA-2S are asymptotically exact for Limiting Regime 1, i.e. Assumptions 1 and 2 in Section I-A are removed, and Assumption 3 becomes true (first scalable asymptotic exactness results for NH-OLSs with non-Poisson fresh requests).
  - The first fairly accurate yet computationally efficient approximation for NH-OLSs with non-Poisson input, namely IESA-2S.

- Overall, the further development of the decomposition methodology for blocking probability evaluation in NH-OLSs, making fundamental contributions in teletraffic modeling.

II. Background and Related Work

A. Mutual Overflow in NH-OLSs

It is well known that non-hierarchical systems with mutual overflow are often more efficient than their hierarchical counterparts. For example, electromechanical switching systems in early telephony were often designed as "slipped gradings" [19], [20], so that each outlet would be the first choice for a given set of incoming calls, the second choice for another set of incoming calls, and so on. Non-hierarchical architectures are also used to control the routing of calls *between* switches, with AT&T's implementation of dynamic non-hierarchical routing providing significant cost savings compared to the previous hierarchical network [40], and real-time network routing [41] providing further cost savings in addition to improved quality of service. Mutual overflow has also been shown to provide advantages in packet-switching networks [42], manufacturing [43], and emergency healthcare [8], and arises naturally in systems with spatial considerations, such as cellular networks [1]–[4], file-sharing networks [7], and emergency vehicular dispatch networks [9], [10], where server preference depends on proximity.

B. Non-Poisson Arrivals in OLSs

The non-Poisson nature of arrivals is an important property of packet-switched telecommunications networks. Both wired and wireless networks have evolved over time to carry more and more packet-based traffic; in fact, in modern telecommunication networks, even voice data is now packetized (i.e. voice over LTE [44]). As packet traffic is burstier than Poisson [45], it is important that blocking probability evaluation methods for OLSs take the burstiness of the offered traffic into account. In other words, the assumption of Poisson arrivals, while suitable for circuit-switched networks (where arrivals are call requests), is no longer applicable to packet-based telecommunication systems and networks. On the other hand, in optical networks, optical transmissions generally consist of many consolidated IP packets, which has the effect of smoothing out arrivals so that the arrival process of optical bursts is generally smoother than Poisson [39].

Non-Poisson arrival processes are also useful in applying overflow loss models to applications outside of telecommunications. For example, while Poisson processes are well-suited for modeling arrivals of emergency patients to intensive care units [46]–[48], this was found not to apply to patients undergoing scheduled elective operations [47], [49]. Non-Poisson arrival processes are also observed for tasks arriving to a cloud computing service (for example, a Poisson *batch* process is considered in [50]) and video-on-demand sessions [51].

C. Analytical Approximation Methods for Blocking Probability Evaluation in NH-OLSs

One analytical approximation method for blocking probability evaluation in NH-OLSs assumes complete homogeneity, thus approximating the system as Erlang's Ideal Grading (EIG) [18]. EIG is an NH-OLS that assumes that all requests have the *same* exponential service time distribution, arrive according to a Poisson process, and that each request may attempt the *same* maximum number of servers and may attempt these servers in random order. Erlang's method for computing the blocking probability of EIG is known as *Erlang's interconnection formula* (EIF) [18]. Unfortunately, the simplicity of the EIG model means that EIF is ill-suited for evaluating blocking probability in systems with heterogeneous loading, heterogeneous routing and/or non-Poisson arrivals of fresh requests.

Another method, and the focus of this paper, is decomposition. Decomposition is a method for approximating blocking probability in NH-OLSs in a scalable manner by decomposing a *system model* of an NH-OLS into independent subsystems. A *node model* is then applied to each subsystem to estimate the blocking probability of each individual subsystem.

Traditionally, decomposition-based approximation methods applied decomposition directly to the "true" system model [15], [26], [27]. This is known to produce large approximation errors when applied to NH-OLSs [31], due to the state dependencies between the server groups being ignored. Therefore, *surrogate* system models have been developed in the literature to capture these dependencies in a way that is preserved when decomposition is applied. These include the the preemptive priority (PP) system model [5], [31], [33] and the information exchange surrogate (IES) system model [4], [8], [32], [35], [36], described in detail in Section III-B. As the IES system model was shown in [32] to outperform the PP system model, we shall focus on the IES system model in this paper.

Once a system has been decomposed into independent subsystems, a node model is required to describe the arrival process of fresh and overflow requests to each subsystem. As the actual arrival process of overflow requests to a subsystem is very complex, node models generally seek to approximate this process using a simpler process. For example, in EFPA, the "true" system model is paired with an Erlang B node model [28], in which both fresh and overflow requests to each server group are combined and modeled as a single Poisson process. The Erlang B node model has also been applied to the PP system model in [5], [31], and [33] and the IES system model in [4], [32], and [36]. Other node models model the combined arrival process to a server group as an IPP [37] or as the overflow from an Erlang B queue [52], while others separate fresh and overflow traffic into multiple arrival streams [34], [53], [54]. In this paper, we consider a scalable two-stream node model for non-Poisson input traffic as described in Section V-A. In addition, node models for processor-sharing queues are considered for the true system model in [7] and for the IES system model in [35].

III. System Models

A. True System Model

We consider an NH-OLS with G server groups and G requests types, so that requests of type g, $g = 1, 2, \ldots, G$, will always attempt group g first, then up to $k_g - 1$ server groups

in random order, selected at random from the remaining $G-1$ groups. Requests of type g, $g = 1, 2, \ldots, G$, are blocked and cleared from the system if and only if all k_g server groups attempted are fully occupied, i.e. no idle servers are available in any of these k_g groups. The number of servers in group g, $g = 1, 2, \ldots, G$, is denoted as N_g. The arrival process of fresh requests to each server group is independent of that to the other server groups, and the service time of each request is exponentially distributed with unit mean.

To simplify our analysis, we will let $N_g = N$ and $k_g = k$ for all $g = 1, 2, \ldots, G$, unless otherwise stated. Heterogeneous cases where N_g and k_g take on different values for different values of g are considered in Section VII-C.

B. IES System Model

The IES system model [4], [32], [35], [36], [55] is designed to address the independence assumption inherent in decomposition-based approximations based on the true system model. While similar to the true system model, the IES system model is based on applying an *information exchange mechanism* to the NH-OLS based on each request's estimate of the number of busy server groups in the system. Incoming requests to the system can thus learn about the state of the system as they overflow from one server group to the next. This information is used to control the overflow behavior of requests offered to the system. In this way, state dependencies between server groups are encoded into the requests themselves, and are preserved when the system is decomposed into independent server groups. For more details regarding the rationale and general design principles of the IES system model, see [4], [36].

Under the information exchange mechanism of the IES system model, each request in the system has two attributes: Δ, the set of visited server groups, and Ω, the estimated number of fully occupied server groups in the system (in layered systems, such as those studied in [36], each layer is treated as if it were a separate system). The value Ω forms the basis of information exchange: when a request encounters a fully occupied server group, it will swap Ω values with the most senior (highest Ω) request in service if and only if the incoming request is junior to (has a lower Ω value than) that request. Fresh requests to the system always start with $\Delta = \emptyset$ and $\Omega = 0$.

Consider an overflowed request, such that the maximum number of server groups which may be attempted is k. Suppose this request has just overflowed from an arbitrary server group with a given Δ and Ω value, such that $|\Delta| = n$. By assuming full independence between Δ and Ω, the probability that all remaining $k-n$ attempts will all encounter fully occupied server groups is estimated as

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

In other words, the remaining $k-n$ attempts and the identity of the Ω fully occupied server groups are assumed to be random and independent of the set of previously visited server groups. This estimate is used to control the actual behavior of overflowing requests in the surrogate system: with probability $P_{\Omega,n}$, the request will *abandon* the system without attempting the remaining $k-n$ server groups. As a result, the blocking probability of the IES system model is generally slightly higher than that of the true system model. This is partially offset by the approximation error caused by the decomposition stage; due to the preservation of dependency information, this error is much reduced compared to decomposition of the true system model, as depicted in Fig. 1.

The IES mechanism creates a special hierarchical traffic structure where all requests with an Ω value of j are unaffected by the existence of any requests with an Ω value greater than j. We will use the term "level j" to refer to an OLS under the IES mechanism where all requests with an Ω value greater than j are removed from consideration. As a result of the hierarchical traffic nature of the information exchange mechanism, decomposition-based approximation methods based on the IES model have closed-form solutions when applied to NH-OLSs. Also, the information exchange mechanism is independent of the chosen node model. For example, [32] and [35] use the Erlang B node model and a processor-sharing model, respectively.

IV. Approximation Methods for NH-OLSs With Poisson Arrivals of Fresh Requests

In this section we derive formulas for two existing decomposition-based approximation methods for NH-OLSs with Poisson arrivals of fresh requests, namely EFPA and IESA. For Poisson input traffic, we use an Erlang B node model, such that the *combined* arrival process of fresh and overflow traffic is treated as a single Poisson process. Let λ denote the offered load to each server group composed of fresh requests. Theoretical results are provided in Section VI-A for the condition where the arrival process of *fresh* requests to each server group is Poisson.

A. EFPA

EFPA combines the true system model with the Erlang B node model. Let $\tilde{a}_n^{\text{EFPA}}$ denote the offered load to each server group composed of requests that have overflowed n times in the system,

$$A^{\text{EFPA}} = \sum_{n=0}^{k-1} \tilde{a}_n^{\text{EFPA}} \tag{2}$$

denote the total offered load to each server group, and

$$b^{\text{EFPA}} = E\left(A^{\text{EFPA}}, N\right) \tag{3}$$

denote the blocking probability of each server group, where $E(A, N)$ denotes the Erlang B formula [28] with A Erlangs of traffic and N servers. Due to the assumption that the states of the server groups are mutually independent, we obtain

$$\tilde{a}_n^{\text{EFPA}} = \begin{cases} \lambda, & n = 0 \\ \tilde{a}_{n-1}^{\text{EFPA}} b, & 1 \le n < k. \end{cases} \tag{4}$$

TABLE II
TABLE OF NOTATIONS FOR IESA

Symbol	Definition	Index range (j)	Index range (n)
λ	Offered load to each server group composed of fresh requests		
$a_{j,n}^{\text{I2S}}$	Offered load to each server group composed of requests with $\lvert\Delta\rvert = n$ and $\Omega = j$	$n \ldots G-1$	$0 \ldots k-1$
$\tilde{a}_{j,n}^{\text{I2S}}$	Offered load to each server group composed of requests with $\lvert\Delta\rvert = n$ and $\Omega = n, n+1, \ldots, j$	$n \ldots G-1$	$0 \ldots k-1$
$w_{j,n}^{\text{I2S}}$	Mean of overflow traffic from each server group composed of requests with $\lvert\Delta\rvert = n$ and $\Omega = j$	$n \ldots G$	$1 \ldots k$
a_j^{I2S}	Offered load to each server group composed of requests with $\Omega = j$	$0 \ldots G-1$	
A_j^{I2S}	Offered load to each server group composed of requests with $\Omega \le j$	$0 \ldots G-1$	
b_j^{I2S}	Blocking probability of requests at level j of the IESA hierarchy	$0 \ldots G-1$	

The above equations form a fixed-point system for which b can be solved using fixed-point iteration [56]. Finally, the overall blocking probability of the system is

$$B^{\text{EFPA}} = \left(b^{\text{EFPA}}\right)^k = 1 - \frac{A^{\text{EFPA}}\left(1 - b^{\text{EFPA}}\right)}{\lambda}. \tag{5}$$

Equation (5) uses the fact that $A^{\text{EFPA}}\left(1 - b^{\text{EFPA}}\right)$ is the carried load of each server group.

B. IESA

IESA combines the IES system model with the Erlang B node model. We define our notation in accordance with Table II. For simplicity, all variables with indices out of range are defined to be equal to zero. By definition:

$$a_{j,n}^{\text{IESA}} = \begin{cases} \lambda, & n = j = 0 \\ 0, & (n = 0) \wedge (j > 0) \\ w_{j,n}^{\text{IESA}}\left(1 - P_{j,n}\right), & 1 \le n < k \end{cases}$$

$$\tilde{a}_{j,n}^{\text{IESA}} = \sum_{i=n}^{j} a_{i,n}^{\text{IESA}} \tag{6}$$

$$a_j^{\text{IESA}} = \sum_{n=0}^{j} a_{j,n}^{\text{IESA}}$$

$$A_j^{\text{IESA}} = \sum_{n=0}^{j} \tilde{a}_{j,n}^{\text{IESA}} = \sum_{i=0}^{j} a_i^{\text{IESA}}, \tag{7}$$

where $P_{j,n}$ is defined according to (1). Applying the node model, we obtain:

$$b_j^{\text{IESA}} = E\left(A_j^{\text{IESA}}, N\right). \tag{8}$$

To obtain $w_{j,n}^{\text{IESA}}$ for $n > 1$, we observe that there are two ways for a request to overflow from a server group as an (n, j)-request, i.e. with $\lvert\Delta\rvert = n$ and $\Omega = j$:

1) A $(n-1, j-1)$-request arriving at a server group finds, with probability b_{j-1}^{IESA}, that all servers are busy serving requests with congestion estimates of $j-1$ or less, meaning that no information exchange occurs and the incoming request overflows with an Ω value of j.
2) A $(n-1, i)$-request arriving at a server group, $i \le j-2$, finds, with probability $b_{j-1}^{\text{IESA}} - b_{j-2}^{\text{IESA}}$, that all servers are busy, with the most senior request in service having a congestion estimates of exactly $j-1$. The two requests swap Ω values, so that the request in service obtains a new Ω value of i and the incoming request overflows with an Ω value of j.

We thus obtain:

$$\begin{aligned} w_{j,n}^{\text{IESA}} &= a_{j-1,n-1}^{\text{IESA}} b_{j-1}^{\text{IESA}} + \tilde{a}_{j-2,n-1}^{\text{IESA}}\left(b_{j-1}^{\text{IESA}} - b_{j-2}^{\text{IESA}}\right) \\ &= \tilde{a}_{j-1,n-1}^{\text{IESA}} b_{j-1}^{\text{IESA}} - \tilde{a}_{j-2,n-1}^{\text{IESA}} b_{j-2}^{\text{IESA}} \end{aligned} \tag{9}$$

for all $1 \le n < k$ and $n \le j < G$, where $\tilde{a}_{-1,n}^{\text{IESA}} = 0$ for all n. The overall system blocking probability is

$$\begin{aligned} B^{\text{IESA}} &= \frac{\sum_{n=1}^{k}\sum_{j=n}^{G} w_{j,n}^{\text{IESA}} P_{j,n}}{\lambda} \\ &= 1 - \frac{A_{G-1}^{\text{IESA}}\left(1 - b_{G-1}^{\text{IESA}}\right)}{\lambda}, \end{aligned} \tag{10}$$

where $A_{G-1}^{\text{IESA}}\left(1 - b_{G-1}^{\text{IESA}}\right)$ is the carried traffic per server group at the final (i.e. $(G-1)^{\text{th}}$) level of the IESA hierarchy.

V. APPROXIMATION METHODS FOR NH-OLSs WITH NON-POISSON INPUT

A. Two-Stream Node Model

We model the arrival process to each server group using two separate arrival streams: a Markov process (e.g. IPP or Engset process) representing the arrival stream of all fresh requests, and a Poisson process representing the arrival stream of all overflow requests. We define $\mathcal{F}$ as the arrival process of fresh requests to each server group and $p_i\left(\mathcal{F}, a_p, N\right)$ as the blocking probability of a server group for fresh ($i = 1$) or overflow ($i = 2$) requests, for a server group with N servers offered process $\mathcal{F}$ of fresh traffic and Poisson overflow traffic with intensity a_p. Let λ denote the mean intensity of process $\mathcal{F}$. In Section VI, we show that as $G \to \infty$ with k fixed, the arrival process of overflow requests to each server group in the NH-OLS becomes Poisson, meaning that our modeling of overflow traffic as Poisson is asymptotically exact.

In the following subsections, we derive two new decomposition-based approximation methods for our NH-OLS model for non-Poisson input traffic, called EFPA-2S and IESA-2S, by combining the new two-stream node model with the true system model and the IES system model, respectively. In Section VI-B, we present theoretical results for our NH-OLS model for the condition where the arrival process of fresh requests is non-Poisson.

B. EFPA-2S

Recall that EFPA-2S combines the true system model with the two-stream node model. Let b^{E2S} denote the congestion probability of *overflow* requests (i.e. requests with one or more previous overflows) and $\hat{b}^{\mathrm{E2S}}$ denote the congestion probability of fresh requests. Then

$$\hat{b}^{\mathrm{E2S}} = p_1\left(\mathcal{F}, A^{\mathrm{E2S}} - \lambda, N\right) \tag{11}$$

$$b^{\mathrm{E2S}} = p_2\left(\mathcal{F}, A^{\mathrm{E2S}} - \lambda, N\right) \tag{12}$$

$$A^{\mathrm{E2S}} = \sum_{n=0}^{k-1} \tilde{a}_n^{\mathrm{E2S}} \tag{13}$$

$$\tilde{a}_n^{\mathrm{E2S}} = \begin{cases} \lambda, & n = 0 \\ \lambda \hat{b}^{\mathrm{E2S}}, & n = 1 \\ \tilde{a}_{n-1}^{\mathrm{E2S}} b^{\mathrm{E2S}}, & 2 \le n < k \end{cases} \tag{14}$$

$$\begin{aligned} B^{\mathrm{E2S}} &= \hat{b}^{\mathrm{E2S}}\left(b^{\mathrm{E2S}}\right)^{k-1} \\ &= 1 - \frac{\lambda\left(1-\hat{b}^{\mathrm{E2S}}\right) + \left(A^{\mathrm{E2S}} - \lambda\right)\left(1 - b^{\mathrm{E2S}}\right)}{\lambda}, \end{aligned} \tag{15}$$

where $\tilde{a}_n^{\mathrm{E2S}}$ and A_n^{E2S} correspond to their counterparts for EFPA. In (15), $\lambda\left(1-\hat{b}^{\mathrm{E2S}}\right)$ denotes the carried load for each server group of fresh requests, and $\left(A^{\mathrm{E2S}} - \lambda\right)\left(1 - b^{\mathrm{E2S}}\right)$ denotes the carried load for each server group of overflow requests.

C. IESA-2S

Recall that IESA-2S combines the IES system model with the two-stream node model. Let b_j^{I2S} denote the congestion probability of *overflow* requests (i.e. requests with one or more previous overflows) and $\hat{b}_j^{\mathrm{I2S}}$ denote the congestion probability of fresh requests, at level j of the IES hierarchy. We define the remainder of our notation in accordance with Table II, changing the superscript "IESA" to "I2S". We obtain:

$$a_{j,n}^{\mathrm{I2S}} = \begin{cases} \lambda, & n = j = 0 \\ 0, & (n = 0) \wedge (j > 0) \\ w_{j,n}^{\mathrm{I2S}}\left(1 - P_{j,n}\right) & 1 \le n < k, \end{cases}$$

$$\tilde{a}_{j,n}^{\mathrm{I2S}} = \sum_{i=n}^{j} a_{i,n}^{\mathrm{I2S}}$$

$$a_j^{\mathrm{I2S}} = \sum_{n=0}^{j} a_{j,n}^{\mathrm{I2S}}$$

$$A_j^{\mathrm{I2S}} = \sum_{n=0}^{j} \tilde{a}_{j,n}^{\mathrm{I2S}} = \sum_{i=0}^{j} a_i^{\mathrm{I2S}}.$$

Applying the node model, we obtain

$$\begin{aligned} \hat{b}_j^{\mathrm{I2S}} &= p_1\left(\mathcal{F}, A_j^{\mathrm{I2S}} - \lambda, N\right) \\ b_j^{\mathrm{I2S}} &= p_2\left(\mathcal{F}, A_j^{\mathrm{I2S}} - \lambda, N\right). \end{aligned}$$

Using a similar argument as for (9), we obtain:

$$\begin{aligned} w_{j,n}^{\mathrm{I2S}} &= a_{j-1,n-1}^{\mathrm{I2S}} b_{j-1}^{\mathrm{I2S}} + \tilde{a}_{j-2,n-1}^{\mathrm{I2S}}\left(b_{j-1}^{\mathrm{I2S}} - b_{j-2}^{\mathrm{I2S}}\right) \\ &= \tilde{a}_{j-1,n-1}^{\mathrm{I2S}} b_{j-1}^{\mathrm{I2S}} - \tilde{a}_{j-2,n-1}^{\mathrm{I2S}} b_{j-2}^{\mathrm{I2S}} \end{aligned} \tag{16}$$

for all $2 \le n < k$ and $n \le j < G$. For $n = 1$, we obtain $w_{j,1}^{\mathrm{I2S}} = \tilde{a}_{j,0}^{\mathrm{I2S}} \hat{b}^{\mathrm{I2S}} = \lambda \hat{b}^{\mathrm{I2S}}$. Finally, the overall system blocking probability is

$$\begin{aligned} B^{\mathrm{I2S}} &= \frac{\sum_{n=1}^{k}\sum_{j=n}^{G} w_{j,n}^{\mathrm{I2S}} P_{j,n}}{\lambda} \\ &= 1 - \frac{\lambda\left(1-\hat{b}_{G-1}^{\mathrm{I2S}}\right) + \left(A_{G-1}^{\mathrm{I2S}} - \lambda\right)\left(1 - b_{G-1}^{\mathrm{I2S}}\right)}{\lambda}, \end{aligned} \tag{17}$$

where $\lambda\left(1 - \hat{b}_{G-1}^{\mathrm{I2S}}\right)$ is the carried traffic per server group for fresh requests, and $\left(A_{G-1}^{\mathrm{I2S}} - \lambda\right)\left(1 - b_{G-1}^{\mathrm{I2S}}\right)$ is the same for overflow requests, at the final (i.e. $(G-1)^{\mathrm{th}}$) level of the IES hierarchy.

D. A Note on Arrival Process $\mathcal{F}$

Note that EFPA-2S and IESA-2S are agnostic to the nature of the arrival process $\mathcal{F}$ of fresh requests to each server group and the implementation of the functions $p_1\left(\mathcal{F}, a_p, N\right)$ and $p_2\left(\mathcal{F}, a_p, N\right)$. In other words, *any* arrival process for fresh requests may be used in these two approximation methods, as long as $p_1\left(\mathcal{F}, a_p, N\right)$ and $p_2\left(\mathcal{F}, a_p, N\right)$ are computable.

VI. Theoretical Results

In the limit as $G \to \infty$ (with k fixed), two important phenomena appear. Firstly, the two-stream model becomes asymptotically exact, as long as the arrival process of fresh requests to each server group is exactly modeled. Secondly, the arrival process of requests to each server group becomes independent of that to the other groups.

We shall provide an intuitive explanation of these phenomena as follows. Firstly, any possible dependencies between two overflow arrivals to a server group are reduced by a factor of infinity as $G \to \infty$. Therefore, by the central limit theorem, the arrival process to each server group tends to Poisson as $G \to \infty$ and the two-stream model exactly models the behavior of overflow traffic in the limiting case. Secondly, at any given moment in time, the next requests to two different server groups almost surely come from different sources (i.e. server groups of the original attempts) as $G \to \infty$, and therefore the interarrival distributions of these two server groups are independent, i.e. the arrival processes of the two groups are independent.

We shall formalize these arguments below:

Lemma 1: As $G \to \infty$ with k fixed, the arrival process of overflow requests to each server group becomes Poisson.

Proof: Number the server groups arbitrarily from 1 to G, and let t denote some arbitrary length of time, $0 < t < \infty$. For $i = 2, 3, \ldots G$, let $N_i\,(G)$ denote the number of arrivals in an interval of length Gt to server group 1 of overflow requests originating from server group i, i.e. group i was the initial service attempt. Since, as $G \to \infty$, the probability that any two finite and disjoint sets of arrivals share *any* previously attempted server groups in common tends to zero, the values $N_2\,(G)\,, N_3\,(G)\,, \ldots$ and $N_G\,(G)$ become mutually independent. Finally, the total number of arrivals of overflow requests to Group 1 in that interval is the sum $\sum_{i=2}^{G}\,(N_i\,(G)\,/G)$, which is Poisson distributed by the central limit theorem as $G \to \infty$. ∎

Corollary 1: The two-stream node model is asymptotically exact as $G \to \infty$ with k fixed as long as the arrival process of fresh requests to each server group is exactly modeled.

Lemma 2: As $G \to \infty$ with k fixed, the arrival process of requests to each server group becomes independent of that to any other server group.

Proof: Consider the system at time t and let $f_1\,(t)$ and $f_2\,(t)$ be the time until the next arrival to each of two arbitrary server groups, labeled 1 and 2, respectively. Since a request that has attempted server group 1 will attempt server group 2 with probability zero as $G \to \infty$, the two arrivals to server groups 1 and 2, respectively, almost surely come from two different requests. The probability that these two requests come from the same source (i.e. server groups of the original attempts) tends to zero as $G \to \infty$, and the probability that the two requests have attempted *any* server groups in common also tends to zero as $G \to \infty$. Therefore, $f_1\,(t)$ and $f_2\,(t)$ are independent and the arrival processes of server groups 1 and 2 are also independent. ∎

A. Theoretical Results for EFPA and IESA for Poisson Input Traffic

Proposition 1: IESA converges to EFPA as $G \to \infty$.

Proof: As $G \to \infty$, we obtain

$$P_{j,n} = \begin{cases} 0, & (n < k) \vee (j < G) \\ 1, & \text{otherwise.} \end{cases}$$

Therefore, $a_{j,n}^{\text{IESA}} = w_{j,n}^{\text{IESA}}$ for all $1 \le n < k$ and all $0 \le j < G$. Applying (6) and (9),

$$\begin{aligned} \tilde{a}_{j,n}^{\text{IESA}} &= \sum_{i=n}^{j} w_{j,n}^{\text{IESA}} \\ &= \sum_{i=n}^{j} \left(\tilde{a}_{i-1,n-1}^{\text{IESA}} b_{j-1}^{\text{IESA}} - \tilde{a}_{i-2,n-1}^{\text{IESA}} b_{j-2}^{\text{IESA}} \right) \\ &= \tilde{a}_{j-1,n-1}^{\text{IESA}} b_{j-1}^{\text{IESA}} - \tilde{a}_{n-2,n-1}^{\text{IESA}} b_{j-1}^{\text{IESA}} \\ &= \tilde{a}_{j-1,n-1}^{\text{IESA}} b_{j-1}^{\text{IESA}}. \end{aligned}$$

Thus for all $0 \le n < k$ and $n \le j < G$, we obtain

$$\tilde{a}_{j,n} = \begin{cases} \lambda, & n = 0 \\ \tilde{a}_{j-1,n-1}^{\text{IESA}} b_{j-1}^{\text{IESA}}, & 1 \le n < k. \end{cases} \tag{18}$$

Equations (7), (8), and (18) are equivalent to (2)–(4), with the exception of the additional subscript j. Therefore, as $G \to \infty$, $\lim_{j\to\infty} \tilde{a}_{j,n}^{\text{IESA}}$ approaches $\tilde{a}_n^{\text{EFPA}}$, $\lim_{j\to\infty} A_j^{\text{IESA}}$ approaches A^{EFPA}, $\lim_{j\to\infty} \hat{b}_j^{\text{IESA}}$ approaches b^{EFPA}, and $\lim_{j\to\infty} b_j^{\text{IESA}}$ approaches b^{EFPA}. Therefore, as $G \to \infty$, B^{IESA} (as computed via (10)) approaches B^{EFPA} (as computed by (3)). This completes the proof. ∎

Intuitively, in the limit as $G \to \infty$, requests in the IES surrogate system model never abandon the system unless $|\Delta| = k$, i.e. $P_{\Omega,n} = 0$ for all $\Omega < \infty$ and all $n = |\Delta| < k$. In other words, the IES surrogate system model converges to the true system model in the limit as $G \to \infty$, and therefore IESA converges to EFPA.

Proposition 2: For Poisson arrivals of fresh request to each server group, EFPA and IESA are both asymptotically exact as $G \to \infty$ with k fixed.

Proof: The arrival process of fresh requests to each server group is defined to be Poisson and independent of that to any other server group. By Lemmas 1 and 2, the same applies to the arrival process of overflow requests when $G \to \infty$. Thus all three EFPA assumptions, as listed in Section I-A, are true in the limiting case. Therefore, EFPA is asymptotically exact. By Proposition 1, IESA is also asymptotically exact. ∎

Whereas the previous lemmas and propositions consider the case of fixed k, we now consider the case of k increasing along with G.

Proposition 3: For Poisson arrivals of fresh requests to each server group and $k = G$, IESA is equivalent to a previous approximation known as the Overflow Priority Classification Approximation [31].

For a description of OPCA, a derivation of formulas for our NH-OLS and a proof of Proposition 3, see the Appendix. Based on theoretical results in [33], we provide the following corollaries for Proposition 3:

Corollary 2: For Poisson arrivals of fresh request to each server group, $k = G$, and $N = 1$, $B^{EFPA} \le B^{IESA} \le B$, where B is the actual blocking probability of the system, i.e., IESA is always at least as accurate as EFPA.

Corollary 3: For Poisson arrivals of fresh requests to each server group, critical loading ($\lambda = 1$), $k = G$, and $N = 1$, $1 \le B/B^{IESA} \le \sqrt{2}$.

As the same ratio for EFPA, i.e. B/B^{EFPA}, tends to infinity as $k = G \to \infty$, as proved in [33], we have a limiting regime where the approximation error of IESA is bounded but that of EFPA is not.

B. Theoretical Results for EFPA-2S and IESA-2S for Non-Poisson Input Traffic

Proposition 4: IESA-2S converges to EFPA-2S as $G \to \infty$.

The proof of Proposition 4 is similar to that for Proposition 1 and uses the same intuitive argument: the IES surrogate system model converges to the true model in the limit as $G \to \infty$, and therefore IESA-2S converges to EFPA-2S.

Proposition 5: EFPA-2S and IESA-2S are both asymptotically exact as $G \to \infty$ with k fixed, if the node model exactly

models the arrival process of fresh requests to each server group.

Proof: The arrival process of arrival requests to each server group is defined to be independent of that to any other server group. By Lemmas 1 and 2, the arrival process of overflow requests to each server group is defined to be Poisson and independent of that to any other server group when $G \to \infty$. In other words, the two-stream node model is asymptotically exact as $G \to \infty$, and Assumption 3 of EFPA, as listed in Section I-A, is true in the limiting case, while Assumptions 1 and 2 have been removed. Therefore, EFPA-2S is asymptotically exact as $G \to \infty$. Due to Proposition 4, IESA-2S is also asymptotically exact as $G \to \infty$. ■

VII. Numerical Results

In this section, we consider NH-OLSs with different parameters and compare the accuracy of EFPA, EFPA-2S, IESA, and IESA-2S compared to simulation results. Error bars in each plot represent the 95% confidence intervals of the simulation results, as calculated using Student's t-distribution, but may be too small to see in some cases.

For IPPs, we define λ' as the arrival rate in the on state, γ as the transition rate from the on state to the off state, and ω as the transition rate from the off state to the on state. For Engset processes, we define S as the number of sources and λ' as the arrival rate per Engset source. For all arrival processes, we define λ as the mean arrival rate and z as the peakedness.

A. Numerical Results for Homogeneous NH-OLSs

1) Blocking Probability With Respect to G: We consider an NH-OLS where each request may attempt up to $k = 10$ server groups, with $N = 20$ servers per group. The arrival process of fresh requests to each server group is an IPP with $\lambda' = 26.174$, $\gamma = 4.793$, and $\omega = 10.555$, yielding $\lambda = 18$ and $z = 1.5$ (see [38], [57] on how to compute λ and z from λ', γ, and ω). The blocking probability of the system is shown in Fig. 2 for various values of G, the total number of server groups in the system. The results show numerical support for Propositions 4 and 5. Furthermore, it can be seen that EFPA and EFPA-2S are constant in G. This can be confirmed by observing that the formulas for EFPA and EFPA-2S, in Sections IV-A and V-B, respectively, do not involve the parameter G in any way. This is because decomposition of the true system model assumes that the states of all server groups are mutually independent. Finally, IESA-2S is more accurate than the other three approximations for *all* values of G, and is fairly accurate even for small to moderate values of G despite the node model being designed based on asymptotic behavior, as it is the only approximation to address all three simplifying assumptions of EFPA listed in Section I-A.

The running times for simulation and the four approximation methods are also shown in Fig. 2. The results demonstrate that all four approximations are several orders of magnitude faster than simulation.

Finally, we replace each IPP in the original scenario with an Engset process with $S = 30$ and $\lambda' = 1.5$, yielding $\lambda = 18$ and $z = 0.4$. The results, depicted in Fig. 3, also show support for Propositions 4 and 5. Note that EFPA-2S and IESA-2S produces smaller blocking probability estimates than EFPA and IESA, respectively, when the arrival processes of fresh requests are smooth. With the exception of where the lines for simulation and IESA-2S cross the line for EFPA, IESA-2S is the most accurate of the four approximation methods.

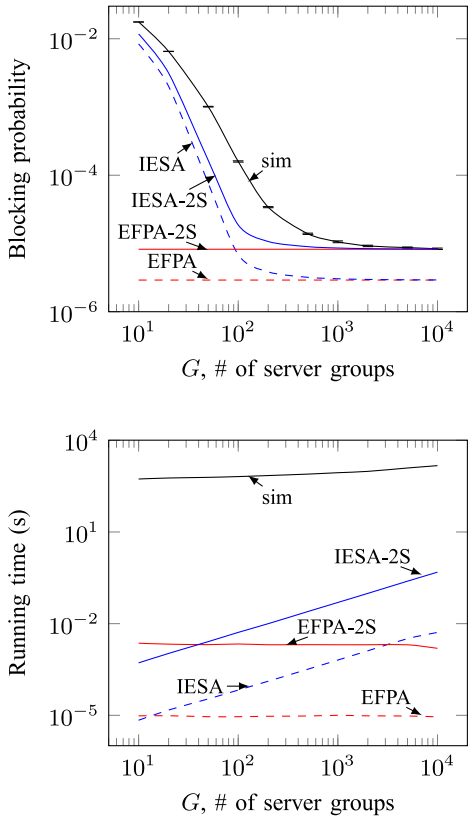


Fig. 2. Blocking probability of an NH-OLS with $k = 10$ and $N = 20$ with respect to G, with fresh requests to each server group forming an IPP with a mean of 18 Erlangs and $z = 1.5$; also the running time for each evaluation method.

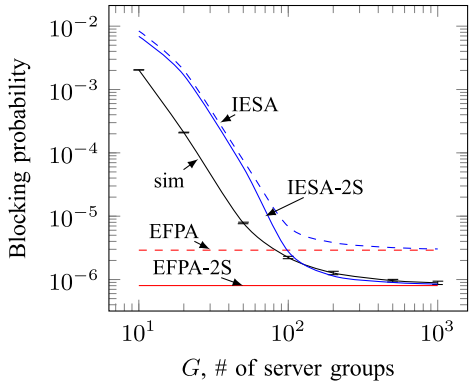


Fig. 3. Blocking probability of an NH-OLS with $k = 10$ and $N = 20$ with respect to G, with fresh requests to each server group forming an Engset process with a mean of 18 Erlangs and $z = 0.4$.

2) Blocking Probability With Respect to k: We consider an NH-OLS with $G = 20$ groups of $N = 20$ servers each. The arrival process of fresh requests to each server group is an IPP with $\lambda' = 26.174$, $\gamma = 4.793$, and $\omega = 10.555$, yielding $\lambda = 18$ and $z = 1.5$. The blocking probability is

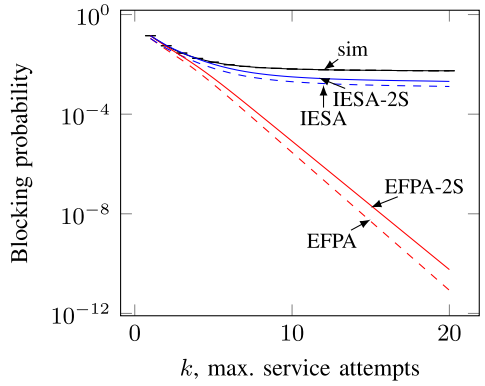


Fig. 4. Blocking probability of an NH-OLS with $G = 20$ and $N = 20$ with respect to k, with fresh requests to each server group forming an IPP with a mean of 18 Erlangs and $z = 1.5$.

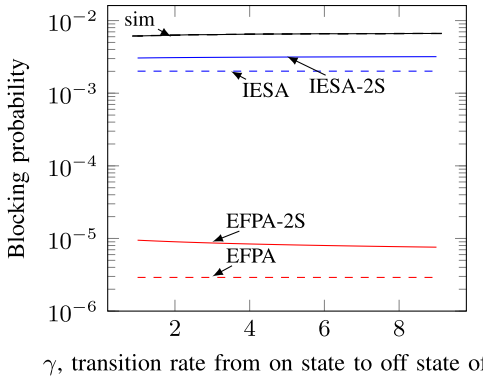


Fig. 5. Blocking probability of an NH-OLS with $G = 20$, $k = 10$, and $N = 20$ with respect to γ, with fresh requests to each server group forming an IPP with a mean of 18 Erlangs and a peakedness of $z = 1.5$.

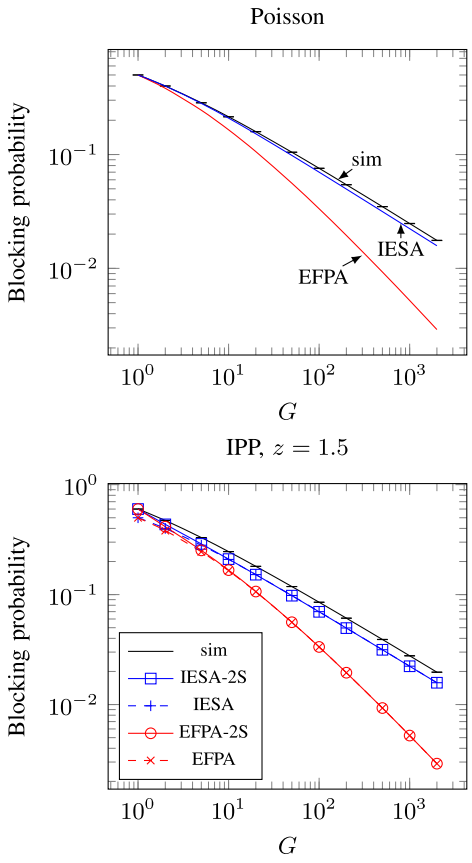


Fig. 6. Blocking probability for an NH-OLS with respect to G, for $G = k$, $N = 1$, and critical loading.

shown in Fig. 4 for various values of k, the maximum number of server groups each request may attempt. The results demonstrate very large approximation errors for EFPA and EFPA-2S as k increases. This is because in these approximations, any error in estimating the blocking probability of a single server group is exponentiated by a factor of k for the entire system. Furthermore, as k increases, the level of mutual overflow in the system also increases, which increases the error caused by assuming independence of the states of each server group in the system. In contrast, IESA and IESA-2S, which take dependencies between server groups into account using the IES surrogate system model, are much more accurate than EFPA and EFPA-2S for large k.

3) Blocking Probability With Respect to the Transition Rates Between the Arrival Process States: We consider an NH-OLS with $G = 20$ groups of $N = 20$ servers each, where each request may attempt up to $k = 10$ server groups. The arrival process of fresh requests to each server group is an IPP with $\omega = \left(\sqrt{\gamma^2 + 146\gamma + 1} - \gamma - 1\right)/2$ and $\lambda' = 18\,(\gamma + \omega)/\omega$, yielding $\lambda = 18$ and $z = 1.5$. The results are shown in Fig. 5 for various values of γ, the transition rate from the on state to the off state. The results demonstrate that the true blocking probability and each of the approximations is not very sensitive to the value of γ, suggesting that λ and z are sufficient to accurately model the system. In all cases, IESA-2S is the most accurate approximation among all approximations considered.

B. Numerical Results for Homogeneous NH-OLSs Under Critical Loading

We consider the special case of critical loading [58]–[60], where the offered load of fresh requests to the system is equal to the system serving capacity. Critical loading is an important scenario since most networks operate in a regime where both offered load and system serving capacity scale upwards in a fixed ratio [59], and critical loading is the maximum loading under such a regime where the blocking probability can be made arbitrary low, e.g. in an Erlang B system.

1) Blocking Probability With Respect to G, Fixed k/G: We consider an NH-OLS under critical loading, with G single-server ($N = 1$) groups and full availability ($k = G$). We first consider the case where the arrival process of fresh requests to each server group is Poisson. Fig. 6 shows the blocking probability of the system for various values of G. The results show numerical support for Corollary 3. On the other hand, it was shown in [33] that $\lim_{G\to\infty} B/B^{\text{EFPA}} = \infty$. In other words, in this limiting regime, the approximation error of IESA is bounded but that of EFPA is not. This demonstrates the importance of having a good surrogate system model that can preserve information on the state dependencies between server groups when decomposition is applied, and

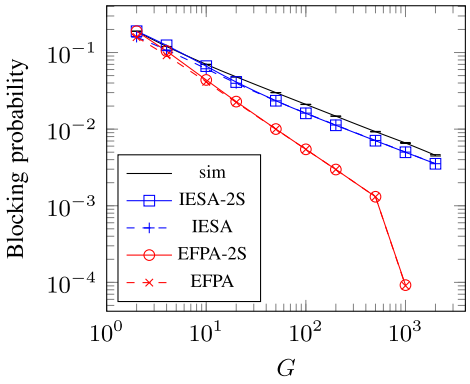


Fig. 7. Blocking probability with respect to G, for $G = 2k$, $N = 20$, and IPP input with critical loading and a peakedness of $z = 1.5$.

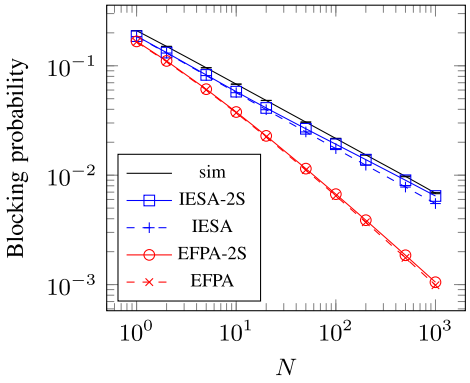


Fig. 8. Blocking probability with respect to N, for $G = 20$, $k = 10$, and IPP input with critical loading and a peakedness of $z = 1.5$.

the cascading effect of increasing k on the error of EFPA as previously demonstrated in Section VII-A2.

We also consider an NH-OLS, again under critical loading and with single-server groups, where the arrival process of fresh requests to each server group is an IPP with $z = 1.5$. The results are shown in Fig. 6 for $k = G$ and Fig. 7 for $k = G/2$ (G even). Although Corollary 3 does not apply in this case, it still appears that the approximation error is bounded for IESA but unbounded for EFPA.

2) Blocking Probability With Respect to N: In this scenario, instead of increasing the system capacity and load by increasing the number of server groups, we increase the size of each server group while keeping the number of groups fixed. We consider an NH-OLS with $G = 20$ single server ($N = 1$) groups, with $k = 10$. The arrival process of fresh requests to each server group is an IPP with $z = 1.5$. The blocking probability is shown in Fig. 8 for various values of N. The results demonstrate that the approximation errors of EFPA and EFPA-2S increase with N, while the errors of IESA and IESA-2S appear to be bounded.

Intuitively, in a system at or near critical loading, the larger the system size, the more sensitive the system is to changes in the offered load. Therefore, in EFPA, any error in estimating the overflow traffic to each server group in large systems leads to increasingly larger errors in the estimation of the overall blocking probability of the system.

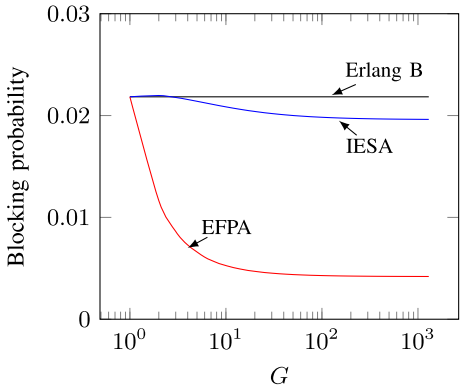


Fig. 9. Blocking probability with respect to G, for $GN = 1296$, $G = k$, and Poisson input with critical loading.

3) Blocking Probability With Respect to G = k, Fixed GN: Among full-availability cases ($G = k$) with a fixed total number of servers (i.e. GN), the case of $N = 1$ is the most challenging as this maximizes k and therefore the error caused by the independence assumption associated with decomposing the true system model (see Section VII-A2 for an intuitive explanation of the effects of increasing k on the accuracy of EFPA). In this subsection, we consider an NH-OLS with G server groups, $GN = 1296$ total servers, and critical loading. The arrival process of fresh requests is a Poisson process. The blocking probability is shown in Fig. 9 for various values of G, demonstrating that the approximation errors of EFPA and IESA are both maximized when $G = 1296$, i.e. when $N = 1$. This again supports our claim that the case of $N = 1$ is the most challenging in terms of conquering the error caused by the independence assumption. As Corollary 3 provides an upper bound on the approximation error of IESA when $N = 1$, we conjecture that this bound also applies to all cases where $N > 1$. In other words:

Conjecture 1: For any NH-OLS with G groups of N servers and full availability ($k = G$), where the arrival rate to each server group of fresh requests is a Poisson process of N Erlangs (critical loading), the ratio of the true blocking probability to the IESA estimate, i.e. B/B^{IESA} is bounded by $\sqrt{2}$.

On the other hand, no such bound exists for EFPA, as proved for $N = 1$ in [33].

C. Numerical Results for Heterogeneous NH-OLSs

In the previous subsections, we assumed that $\lambda_g = \lambda$, $z_g = z$, $k_g = k$, and N_g for all $g = 1, 2, \ldots, G$. In this subsection, we consider cases where some of these assumptions are removed. We consider the following three scenarios for various values of G:

1) $k_g = 10$, $N_g = 20$, $\lambda_g = 18 + 6\,(-1)^g$, and $z_g = 1.5$ for all $g = 1, 2, \ldots, G$;
2) $k_g = 10 + 2\,(-1)^g$, $N_g = 20$, $\lambda_g = 18$, and $z_g = 1.5$ for all $g = 1, 2, \ldots, G$; and
3) $k_g = 10$, $N_g = 20 + 2\,(-1)^g$, $\lambda_g = 18$, and $z_g = 1.5$ for all $g = 1, 2, \ldots, G$.

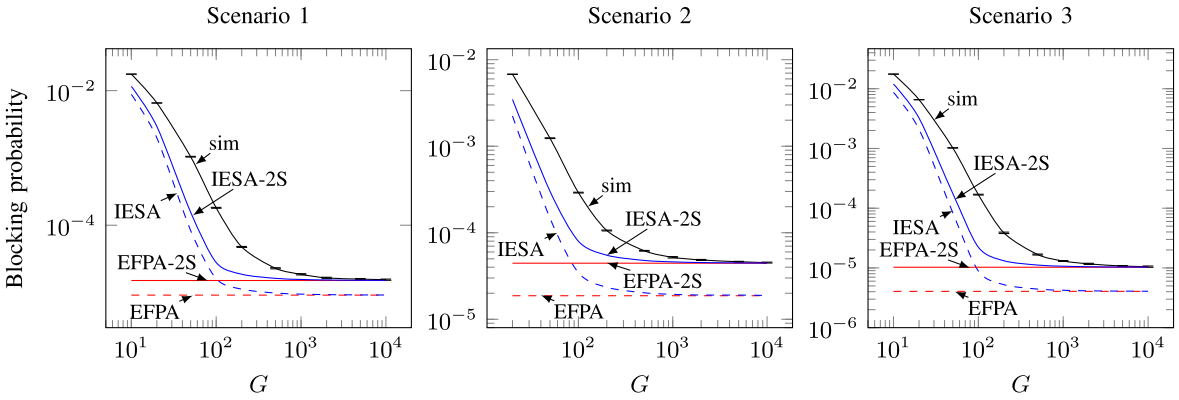


Fig. 10. Blocking probability and logarithmic error with respect to G for Scenarios 1–3 of Section VII-C, regarding heterogenous arrival rates, k_g, and server group sizes, respectively.

The results, shown in Fig. 10, demonstrate that IESA-2S is fairly accurate even for heterogeneous NH-OLSs, especially compared to EFPA, IESA, and EFPA-2S. In particular, the results show that Propositions 4 and 5 also apply to the heterogeneous case, i.e. EFPA-2S and IESA-2S are both asymptotically exact as $G \to \infty$ with k fixed.

VIII. Concluding Remarks

In this paper, we present new developments for the decomposition methodology for blocking probability evaluation in NH-OLSs, making fundamental contributions in teletraffic modeling covering a broad range of NH-OLSs, including those with heterogeneous arrival processes, server group size and/or routing, with both Poisson and non-Poisson arrivals of fresh requests.

We consider a model of an NH-OLS with G server groups and random routing of overflow requests and show that the arrival processes of overflow requests to each server group tend toward independent Poisson processes as $G \to \infty$ with k fixed, provided that the arrival processes of fresh requests to each server group are mutually independent. For NH-OLSs with Poisson input, we use this result (Lemmas 1 and 2) to prove new asymptotic exactness results for EFPA and IESA, which are decomposition methods using the Erlang B node model. Among these are the first scalable asymptotic exactness results for NH-OLSs. We also demonstrate a limiting regime in which the IESA estimate is at least as accurate as the EFPA estimate, and where, under critical loading, the ratio of the true blocking probability to the IESA estimate, i.e. B/B^{IESA}, is bounded above by $\sqrt{2}$, but B/B^{EFPA} is unbounded, showing the benefits of the IESA in capturing mutual state dependencies between server groups which EFPA cannot.

For NH-OLSs with non-Poisson input, we use Lemmas 1 and 2 to develop a new scalable node model which models arrivals to a server group using two traffic streams, one for fresh requests and one for overflow requests, where the arrival stream for overflow requests is modeled as a single Poisson process. We use this new two-stream node model to develop two new decomposition-based approximation methods, namely EFPA-2S and IESA-2S. We show for an NH-OLS with random routing of overflow requests that if the arrival stream of fresh requests to each server group is independent of that to the other groups, and is modeled exactly in the node model for the limiting case, where the overflow traffic to each server group becomes independent Poisson processes, then EFPA-2S and IESA-2S are asymptotically exact as $G \to \infty$ with k fixed, just as EFPA and IESA are asymptotically exact if the arrival process of fresh requests to each server group is Poisson and independent of that to the other groups.

Due to the simplicity of the two-stream node model, both EFPA-2S and IESA-2S are computationally-efficient, provided that the arrival process of fresh requests to each server group can be modeled as a simple process, e.g. IPP or Engset. Furthermore, numerical results show that IESA-2S is generally quite accurate for NH-OLSs with IPP or Engset input even when the number of server groups in the system is limited, especially compared to EFPA, EFPA-2S, and IESA. We conclude that IESA-2S is the first computationally efficient and fairly accurate approximation with asymptotic exactness properties for NH-OLSs with both mutual overflow and non-Poisson input.

The node model developed in this paper can be applied to more complex versions of IESA or EFPA for general NH-OLSs. Note that IESA has been used in the literature to approximate blocking probability in cellular networks [4] and ICU networks [8], while EFPA has been applied to each of the examples in Table I. Applying IESA-2S to applications originally using IESA is simply a matter of replacing the node model (e.g. an Erlang B queueing model) by the new two-stream node model. On the other hand, applying IESA-2S to applications originally using EFPA is simply a matter of applying IESA to replace EFPA but with the new two-stream node model instead of the original single-stream node model.

Appendix
OPCA

A. PP System Model

In the PP system model [31], requests form an hierarchy based on the number of previously attempted server groups.

Each request carries a parameter Δ containing the set of attempted server groups by that request. As in the true model, an incoming request to server group g, $g = 1, 2, \ldots, G$, is served if there is at least one idle server. However, if the server group is full, the request compares its Δ parameter, which we denote Δ_1, with that of the most senior (i.e. highest $|\Delta|$) request in service, which we denote Δ_2. If $|\Delta_1| \geq |\Delta_2|$, then the incoming request overflows normally with a new Δ parameter of $\Delta_1 \cup \{g\}$. However, if $|\Delta_1| < |\Delta_2|$, the incoming request *replaces* the request in service. The request originally in service then overflows as with a Δ parameter of $\Delta_2 \cup \{g\}$. In other words, requests with a smaller $|\Delta|$ have *preemptive priority* in the PP system model.

As a result of the hierarchical traffic nature of the preemptive priority mechanism model, decomposition-based approximation methods based on this model have closed-form solutions when applied to NH-OLSs. Also, the preemptive priority mechanism is independent of the chosen node model. For example, [31] and [35] use the Erlang B node model and a processor-sharing model, respectively.

B. OPCA

OPCA combines the PP system model with the Erlang B node model. Although OPCA has been shown to be less robust than IESA [31], we use it in Section VI-A to prove various analytical results regarding IESA.

Let the term n-request denote a request with $\Delta = n$. Define:

- λ as the offered load of fresh requests to each server group;
- a_n^{OPCA} as the total offered load to a server group composed of n-requests;
- A_n^{OPCA} as the total offered load to a server group composed of i-requests, $i \leq n$; and
- b_n^{OPCA} as the blocking probability of a server group at level n of the PP system hierarchy, that is, the blocking probability of a server group when only i-requests, $0 \leq n$, are considered.

By definition, $A_n^{\mathrm{OPCA}} = \sum_{i=0}^{n} a_i^{\mathrm{OPCA}}$. Since we assume all traffic, fresh and overflow, is Poisson, we obtain

$$b_n^{\mathrm{OPCA}} = E(A_n^{\mathrm{OPCA}}, N). \tag{19}$$

From the OPCA mechanism, we obtain

$$\begin{aligned} a_n^{\mathrm{OPCA}} &= a_{n-1}^{\mathrm{OPCA}} b_{n-1}^{\mathrm{OPCA}} + A_{n-2}^{\mathrm{OPCA}} (b_{n-1}^{\mathrm{OPCA}} - b_{n-2}^{\mathrm{OPCA}}) \\ &= A_{n-1}^{\mathrm{OPCA}} b_{n-1}^{\mathrm{OPCA}} - A_{n-2}^{\mathrm{OPCA}} b_{n-2}^{\mathrm{OPCA}}, \end{aligned} \tag{20}$$

with base cases $a_0^{\mathrm{OPCA}} = A_0^{\mathrm{OPCA}} = \lambda$ and $a_n^{\mathrm{OPCA}} = A_n^{\mathrm{OPCA}} = b_n^{\mathrm{OPCA}} = 0$ for $n < 0$. The above values can be found iteratively for $n = 0, 1, \ldots, k-1$. Finally, the overall blocking probability of the system is estimated as

$$B^{\mathrm{OPCA}} = 1 - \frac{A_{k-1}^{\mathrm{OPCA}} \left(1 - b_{k-1}^{\mathrm{OPCA}}\right)}{\lambda}, \tag{21}$$

where $A_{k-1}^{\mathrm{OPCA}} \left(1 - b_{k-1}^{\mathrm{OPCA}}\right)$ is the total carried load of the system.

Another way to compute A_n^{OPCA} is to apply (20):

$$\begin{aligned} A_n^{\mathrm{OPCA}} &= \sum_{i=0}^{n} a_i^{\mathrm{OPCA}} \\ &= a_0^{\mathrm{OPCA}} + \sum_{i=1}^{n} \left(A_{n-1}^{\mathrm{OPCA}} b_{n-1}^{\mathrm{OPCA}} - A_{n-2}^{\mathrm{OPCA}} b_{n-2}^{\mathrm{OPCA}} \right) \\ &= A_0^{\mathrm{OPCA}} + A_{n-1}^{\mathrm{OPCA}} b_{n-1}^{\mathrm{OPCA}}, \end{aligned} \tag{22}$$

which can be interpreted as the initial offered load of fresh traffic to each server group summed with the total offered load of overflow traffic to each server group at level n of the PP system hierarchy, which contains all requests with $|\Delta| \leq n$.

C. Proof of Proposition 3: Equivalence of OPCA and IESA for $G = k$

Due to the similarities between (8) and (19), and between (10) and (21), it suffices to show simply that $A_j^{\mathrm{IESA}} = A_n^{\mathrm{OPCA}}$ for all $n = j = 0, 1, \ldots, G-1$. Since $P_{j,n} = 0$ for all $n < G-1$ when $G = k$, we obtain $a_{j,n} = w_{j,n}$ for all $1 \leq n \leq j < G$. Applying (6) and (9), we obtain

$$\begin{aligned} \tilde{a}_{j,n}^{\mathrm{IESA}} &= \sum_{i=n}^{j} a_{j,n}^{\mathrm{IESA}} = \sum_{i=n}^{j} w_{j,n}^{\mathrm{IESA}} \\ &= \sum_{i=n}^{j} \left(\tilde{a}_{i-1,n-1}^{\mathrm{IESA}} b_{j-1}^{\mathrm{IESA}} - \tilde{a}_{i-2,n-1}^{\mathrm{IESA}} b_{j-2}^{\mathrm{IESA}} \right) \\ &= \tilde{a}_{j-1,n-1}^{\mathrm{IESA}} b_{j-1}^{\mathrm{IESA}} - \tilde{a}_{n-2,n-1}^{\mathrm{IESA}} b_{n-1}^{\mathrm{IESA}} \\ &= \tilde{a}_{j-1,n-1}^{\mathrm{IESA}} b_{j-1}^{\mathrm{IESA}} \end{aligned} \tag{23}$$

for all $1 \leq n \leq j < G$. Applying (23) to (7), we obtain

$$\begin{aligned} A_j^{\mathrm{IESA}} &= \sum_{n=0}^{j} \tilde{a}_{j,n}^{\mathrm{IESA}} = \tilde{a}_{0,0}^{\mathrm{IESA}} + \sum_{n=1}^{j} \tilde{a}_{j-1,n-1}^{\mathrm{IESA}} b_{j-1}^{\mathrm{IESA}} \\ &= a_{0,0}^{\mathrm{IESA}} + \left(\sum_{n=0}^{j-1} \tilde{a}_{j,n}^{\mathrm{IESA}} \right) b_{j-1}^{\mathrm{IESA}} \\ &= A_0^{\mathrm{IESA}} + A_{j-1}^{\mathrm{IESA}} b_{j-1}^{\mathrm{IESA}} \end{aligned} \tag{24}$$

for all $1 \leq j < G$. Finally, since $A_0^{\mathrm{IESA}} = A_0^{\mathrm{OPCA}} = \lambda$, we can show by induction that $A_j^{\mathrm{IESA}} = A_n^{\mathrm{OPCA}}$ for all $n = j = 0, 1, \ldots, G-1$, using equations (22) and (24). ■

REFERENCES

[1] B. Eklundh, "Channel utilization and blocking probability in a cellular mobile telephone system with directed retry," *IEEE Trans. Commun.*, vol. COM-34, no. 4, pp. 329–337, Apr. 1986.

[2] D. Everitt, "Traffic capacity of cellular mobile communications systems," *Comput. Netw. ISDN Syst.*, vol. 20, nos. 1–5, pp. 447–454, 1990.

[3] 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.

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

[5] E. W. M. Wong, M. Y. M. Chiu, M. Zukerman, Z. Rosberg, S. Chan, and A. Zalesky, "A novel method for modeling and analysis of distributed video on demand systems," in *Proc. IEEE ICC*, May 2005, pp. 88–92.

[6] 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.

[7] 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.

[8] 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 Manage. Sci.*, to be published, doi: 10.1007/s10729-017-9412-8.

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

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

[11] Y. Tan, Y. Lu, and C. H. Xia, "Provisioning for large scale loss network systems with applications in cloud computing," *ACM SIGMETRICS Perform. Eval. Rev.*, vol. 40, no. 3, pp. 83–85, 2012.

[12] Z. Rosberg, Y. Peng, J. Fu, J. Guo, E. W. M. Wong, and M. Zukerman, "Insensitive job assignment with throughput and energy criteria for processor-sharing server farms," *IEEE/ACM Trans. Netw.*, vol. 22, no. 4, pp. 1257–1270, Aug. 2014.

[13] U. R. Krieger, "Modeling and performance analysis of interconnected servers in a cloud computing system with dynamic load balancing," in *Proc. DCCN*, 2015, pp. 52–60.

[14] J. Fu, B. Moran, E. W. M. Wong, and M. Zukerman, "Asymptotically optimal job assignment for energy-efficient processor-sharing server farms," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 4008–4023, Dec. 2016.

[15] G. Koole and J. Talim, "Exponential approximation of multi-skill call centers architecture," in *Proc. QNETs*, 2000, pp. 23/1–23/10.

[16] M. Schneps-Schneppe and J. Sedols, "Markov models for multi-skill call center," *Int. J. Netw. Commun.*, vol. 2, no. 4, pp. 55–61, 2012.

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

[18] A. K. Erlang, "The application of the theory of probabilities in telephone administration," in *The Life Works A. K. Erlang* (Transactions of the Danish Academy of Technical Sciences), vol. 2, E. Brockmeyer, H. L. Halstrøm, and A. Jensen, Eds. Copenhagen, Denmark: Danish Academy of Technical Sciences, 1948, pp. 201–215.

[19] H. A. Longley, "The efficiency of gradings, Part I—Determination of general formulae–small grading elements," *Post Office Electr. Eng. J.*, vol. 41, pp. 45–49, 1948.

[20] H. A. Longley, "The efficiency of gradings, Part II—Grades of service for straight gradings–Comparative efficiencies of various arrangements," *Post Office Electr. Eng. J.*, vol. 41, pp. 67–72, 1948.

[21] A. Lotze, "History and development of grading theory," *Archiv Elektron. Übertragungstechnik*, vol. 25, nos. 9–10, pp. 402–410, 1971.

[22] A. Hordijk, "Insensitive bounds for performance measures," in *Proc. ITC*, 1988, pp. 1–7.

[23] B. Hennion, "Feedback methods for calls allocation on the crossed traffic routing," in *Proc. ITC*, 1979, pp. 1–3.

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

[25] U. R. Krieger, "Analysis of a loss system with mutual overflow in a Markovian environment," in *Proc. 1st Workshop Numer. Solution Markov Chains*, 1991, pp. 303–328.

[26] 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, Tech. Rep. Memo. MM64-3122-2, 1964.

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

[28] A. K. Erlang, "Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges," in *The Life Works A. K. Erlang* (Transactions of the Danish Academy of Technical Sciences), vol. 2, E. Brockmeyer, H. L. Halstrøm, and A. Jensen, Eds. Copenhagen, Denmark: Danish Academy of Technical Sciences, 1948, pp. 138–155.

[29] A. Girard, *Routing and Dimensioning in Circuit-Switched Networks*. Reading, MA, USA: Addison-Wesley, 1990.

[30] 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.

[31] 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, 2007.

[32] 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. ITC*, Sep. 2013, pp. 1–9.

[33] 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," *Stoch. Models*, vol. 29, no. 2, pp. 149–189, 2013.

[34] A. Kuczura, "Loss systems with mixed renewal and Poisson inputs," *Oper. Res.*, vol. 21, no. 3, pp. 787–795, 1973.

[35] Y.-C. Chan, J. Guo, E. W. M. Wong, and M. Zukerman, "Performance analysis for overflow loss systems of processor-sharing queues," in *Proc. IEEE INFOCOM*, Apr./May 2015, pp. 1409–1417.

[36] 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.

[37] K. S. Meier-Hellstern, "The analysis of a queue arising in overflow models," *IEEE Trans. Commun.*, vol. 37, no. 4, pp. 367–371, Apr. 1989.

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

[39] H. Øverby, "Performance modelling of optical packet switched networks with the Engset traffic model," *Opt. Exp.*, vol. 13, no. 5, pp. 1685–1695, 2005.

[40] G. R. Ash, A. H. Kafker, and K. R. Krishnan, "Intercity dynamic routing architecture and feasibility," in *Proc. ITC*, 1983, pp. 1–7.

[41] G. R. Ash, J.-S. Chen, A. E. Frey, B. D. Huang, C.-K. Lee, and G. L. McDonald, "Real-time network routing in the AT&T network-improved service quality at lower cost," in *Proc. IEEE GLOBECOM*, Dec. 1992, pp. 802–809.

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

[43] S. C. Graves, "Flexibility principles," in *Building Intuition* (International Series in Operations Research & Management Science), vol. 115. Boston, MA, USA: Springer, 2008, ch. 3, pp. 33–49.

[44] M. Poikselkä, H. Holma, J. Hongisto. J. Kallio, and A. Toskala, *Voice Over LTE: VoLTE*. Hoboken, NJ, USA: Wiley, 2012.

[45] V. Paxson and S. Floyd, "Wide area traffic: The failure of Poisson modeling," *IEEE/ACM Trans. Netw.*, vol. 3, no. 3, pp. 226–244, Jun. 1995.

[46] J. C. Lowery, "Multi-hospital validation of critical care simulation model," in *Proc. WSC*, Dec. 1993, pp. 1207–1215.

[47] S.-C. Kim, I. Horowitz, K. K. Young, and T. A. Buckley, "Analysis of capacity management of the intensive care unit in a hospital," *Eur. J. Oper. Res.*, vol. 115, no. 1, pp. 36–46, 1999.

[48] 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, 2008.

[49] S.-C. Kim, I. Horowitz, K. K. Young, and T. A. Buckley, "Flexible bed allocation and performance in the intensive care unit," *J. Oper. Manage.*, vol. 18, no. 4, pp. 427–443, 2000.

[50] H. Khazaei, J. Mišić, and V. B. Mišić, "Performance of cloud centers with high degree of virtualization under batch task arrivals," *IEEE Trans. Parallel Distrib. Syst.*, vol. 24, no. 12, pp. 2429–2438, Dec. 2013.

[51] A. Ali-Eldin, M. Kihl, J. Tordsson, and E. Elmroth, "Analysis and characterization of a video-on-demand service workload," in *Proc. ACM MMSys*, 2015, pp. 189–200.

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

[53] J. Matsumoto and Y. Watanabe, "Individual traffic characteristics queueing systems with multiple Poisson and overflow inputs," *IEEE Trans. Commun.*, vol. COM-33, no. 1, pp. 1–9, Jan. 1985.

[54] 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, 2005.

[55] J. Wu, J. Guo, E. W. M. Wong, and M. Zukerman, "Approximation of blocking probabilities in mobile cellular networks with channel borrowing," in *Proc. IEEE HPSR*, Jul. 2015, pp. 1–6.

[56] 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.

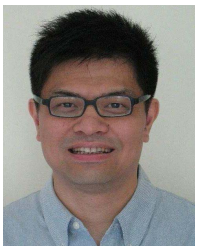
[57] D. L. Jagerman, "Methods in traffic calculations," *Bell Lab. Tech. J.*, vol. 63, no. 7, pp. 1283–1310, 1984.

[58] P. J. Hunt and F. P. Kelly, "On critically loaded loss networks," *Adv. Appl. Probab.*, vol. 21, no. 4, pp. 831–841, 1989.

[59] M. I. Reiman, "Some allocation problems for critically loaded loss systems with independent links," *Perform. Eval.*, vol. 13, no. 1, pp. 17–25, 1991.

[60] J. A. Morrison, "Asymptotic shape of the Erlang capacity region of a critically loaded multiservice shared resource," *SIAM J. Appl. Math.*, vol. 64, no. 1, pp. 1–17, 2003.

Yin-Chi Chan (S'15–M'17) received the B.Math. degree from the University of Waterloo, Waterloo, ON, Canada, in 2010, and the M.Sc. and Ph.D. degrees from the City University of Hong Kong, Hong Kong, in 2011 and 2017, respectively. He is currently a Post-Doctoral Fellow with the Department of Electronic Engineering, City University of Hong Kong. His research interest is currently focused on approximative methods for the performance evaluation, optimization, and design of communications and service systems.



Eric W. M. Wong (S'87–M'90–SM'00) 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, Amherst, MA, USA, in 1994. He is currently an Associate Professor with the Department of Electronic 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, and optical networking.