Performance Evaluation 104 (2016) 1–22

Contents lists available at ScienceDirect

Performance Evaluation

journal homepage: www.elsevier.com/locate/peva

Surrogate models for performance evaluation of multi-skill multi-layer overflow loss systems

Yin-Chi Chan [*], Jun Guo, Eric W.M. Wong, Moshe Zukerman

Department of Electronic Engineering, City University of Hong Kong, 83 Tat Chee Ave., Hong Kong

ARTICLE INFO

Article history:
Received 26 November 2015
Received in revised form 17 June 2016
Accepted 30 June 2016
Available online 7 July 2016

Keywords:
Multi-skill overflow loss system
Blocking probability
Surrogate models
Approximation

ABSTRACT

We consider a model of overflow loss systems in which server groups are arranged into layers, and alternate routing within each layer creates mutual overflow effects, increasing the amount of traffic that can be carried by the system. Such a model has wide applications in communications and service systems. However, the presence of both hierarchical inter-layer overflow and mutual intra-layer overflow makes accurate, robust, yet scalable blocking probability evaluation of such systems a difficult challenge. To address this challenge, we apply and extend the recently developed Information Exchange Surrogate Approximation (IESA) framework to a multi-layer system, adding new surrogate models to the framework and incorporating moment-matching techniques. In contrast to the conventional fixed-point approximation (FPA) approach, which directly decomposes the overflow loss system into independent subsystems with inherent problems of convergence and uniqueness, IESA performs decomposition on a carefully designed surrogate model with guaranteed convergence and uniqueness. Extensive numerical results demonstrate that IESA is consistently more accurate than the conventional FPA approach, showing an improvement in accuracy of several orders of magnitude in many cases. Furthermore, the new extensions to IESA introduced in this paper provide consistent improvements in accuracy relative to the current state-of-the-art of the IESA framework.

1. Introduction

Overflow loss systems are characterized by one or more classes of requests served by a system comprised of multiple server groups, with requests from each class following a prescribed *overflow policy* in seeking an available server [1–4]. They arise naturally in a variety of communications and services systems, for example wireless and cellular networks [5–7], video-on-demand systems [8–10], emergency vehicular dispatch [11–15], and intensive care units [16–18]. Unfortunately, even the simplest overflow loss systems often have no simple analytic expression for the blocking probability of requests [4], since the stationary distribution of an overflow loss system is not of product form. The challenge in practice is thus to find accurate, robust, yet computationally efficient approximation methods.

In particular, many applications of overflow loss systems naturally give rise to multi-layer architectures, yet also allow non-hierarchical intra-layer overflow within each layer. Such a design is motivated by two principles. Firstly, it is well known that in overflow loss systems, it is generally preferable for requests to attempt servers with smaller skill sets before those with larger skill sets (in terms of the number of request types able to be handled by each server) [19]. Secondly, system

[*] Corresponding author.
E-mail addresses: ycchan26-c@my.cityu.edu.hk (Y.-C. Chan), j.guo@cityu.edu.hk (J. Guo), eeewong@cityu.edu.hk (E.W.M. Wong), m.zu@cityu.edu.hk (M. Zukerman).

http://dx.doi.org/10.1016/j.peva.2016.06.007
0166-5316/© 2016 Elsevier B.V. All rights reserved.

efficiency can generally be improved by arranging server groups to form what is known as a *closed chain* [20]. Such closed chains allow temporary overcapacity in any part of the chain to be transferred to handle any temporary capacity shortages in any other part of the chain. Closed chains thus improve the efficiency of each system layer by enhancing the mutual sharing effect between server groups and are closely related to the concept of "entraide" or mutual aid in telephone switching systems [21,22].

The presence of closed chains leads to a phenomenon known as *mutual overflow* [23–25], where congestion on a specific server causes overflow to the other servers, which in turn become congested and yield overflow to the original server. While the classical Fixed Point Approximation (FPA) [26,27] is generally sufficient for approximating blocking in *pure* hierarchical systems, especially when enhanced with moment-matching techniques [28,26,29–31], such methods are generally inadequate when mutual overflow is present [32]. This is because FPA does not capture the mutual dependencies between server groups.

1.1. Addressing mutual overflow

To address mutual overflow in overflow loss systems, the recently developed Information Exchange Surrogate Approximation (IESA) framework [33,7,34] was proposed. IESA is based on applying the underlying methodology of FPA, namely decoupling of a system into multiple independent queues with Poisson input, to a surrogate model of the system that preserves some of the dependency information between server groups when decoupling is applied. As a result, IESA has been shown to provide more accurate and robust results compared to FPA for a number of cases [33]. In fact, IESA appears to be the *first* approximation framework which accurately handles mutual overflow in a heterogeneous system environment, thus addressing a well-known historical problem [35]. In addition, because the surrogate model creates a pure hierarchical traffic structure *within* each layer of the overflow loss system, IESA as applied in this paper does not require the use of fixed-point iteration (unlike FPA when mutual overflow is present), and therefore can be completed in a finite number of steps with guaranteed convergence to a unique solution.

The advantage of IESA over simulation is that IESA provides new insight into and better understanding of the nature of overflow loss systems, with particular focus on the mutual dependency effects between server groups in the same system layer (which are ignored in FPA). In addition, IESA allows fast evaluation of a large number of system configurations, allowing for the optimization of resource allocation in overflow loss systems, including improvements in system design.

1.2. Contributions of this paper

The main contribution of this paper is the extension of the IESA framework to a multi-layer overflow loss system model with intra-layer overflow. We shall use the term "IESA" to refer both to the IESA framework as a whole and to its application in this paper to a multi-layer model. Extensive numerical results demonstrate consistently better accuracy of IESA over FPA, with several orders of magnitude of improvement in many cases.

In addition, we also propose improvements to IESA for capturing the intra-layer dependencies in the overflow loss system. As our new surrogate model is closely related to the previous surrogate model, we shall label the resulting approximation as IESA$^+$. Although the congestion estimates are defined in the same way in both the original and new IESA surrogate models, the way the surrogate model uses these estimates is slightly different. While this paper focuses on the application of IESA to multi-layer overflow loss systems, this improved version of IESA, i.e. IESA$^+$, is equally as applicable to single-layer systems. We shall use the term "true model" to refer to our original overflow loss system model as defined in Section 3, and "IESA surrogate model" and "IESA$^+$ surrogate model" (IESA model and IESA$^+$ model for brevity) to refer to the surrogate models for the IESA and IESA$^+$ approximations, respectively.

Finally, we apply moment matching to FPA, IESA, and IESA$^+$. The moment-matched versions of these approximations are denoted FPAm, IESAm, and IESAm$^+$, respectively. IESAm$^+$ is demonstrated via extensive numerical results to be the most accurate and robust approximation out of all those considered in this paper.

1.3. Applications of multi-layer systems

The multi-layer model in this paper has many applications. One example is cellular networks [36,37,7], where cells can be classified into layers based on coverage area, for example, as macro-cells and micro-cells. The cellular network model is similar to the one studied in this paper, but adds the concepts of call mobility (i.e. handoff of calls between cells) and locality (overflow and handoffs can only occur between adjacent or overlapping cells). Extensions to IESA regarding these two issues were presented in [7], but for a single-layer system only.

Another example is that of content distribution networks (CDNs). For example, the single-layer version [33] of the model considered in this paper is motivated by CDNs for video-on-demand [9]. In a multi-layered CDN design, servers would be divided into origin servers and edge servers, with possible additional layers in between. This allows most popular content in the network to be shifted as close to the end users as possible. In addition, the edge layer of a CDN network may also incorporate peer-to-peer elements [38,10]. As a real-life example of the benefits of multi-layered CDNs, Facebook's cold storage data centers are roughly six times more energy efficient than its regular data centers [39]. The model in [34], to which IESA is applied, is motivated by P2P networks for video-on-demand systems, but is restricted to a single layer.

The multi-layer model can also be applied to Infrastructure as a Service in cloud computing platforms. For example, Amazon Web Services subdivides each of its regions into multiple availability zones, each containing multiple data centers. If a user does not select an availability zone when deploying a virtual machine (VM), Amazon may deploy the VM at any data center in the region. Alternatively, a user may choose to launch a new compute instance in a specific availability zone based on the location of existing storage instances.

Multi-layered models also arise naturally in call centers [30,40], where cross-training costs give rise to differentiation among call center agents. Whereas Franx et al. [30] only considered purely hierarchical call center architectures, the ability of IESA to accurately model mutual overflow within each call center layer allows us to more fully utilize each layer of the call center. Although delay forms a major aspect of call centers in reality, Chevalier and Van den Schrieck [41,40] argue that results obtained with a loss model can be a good proxy for models with waiting.

Finally, layered architectures can be found in hospital management systems. For example, the Hong Kong Hospital Authority currently manages (as of June 2016) 41 public hospitals organized into seven clusters [42]. While it is preferable to serve each patient in his or her preferred cluster, patients may also be referred between clusters for load-balancing reasons or if specialist services are required.

1.4. Organization

The rest of this paper is organized as follows. In Section 2, we discuss existing related work in more detail. In Section 3, we describe the model of multi-layer overflow loss systems considered in this paper. Section 4 illustrates the benefits of layering and mutual overflow that motivate our chosen model. Section 5 gives details of how we apply FPA, IESA, and IESA$^+$ to our chosen model, as well as the corresponding moment-matched versions of these three approximations. The performance of each approximation is compared numerically in Section 6. In Section 6.10, we demonstrate near insensitivity of the blocking probability to the service time distribution, allowing our IESA framework to be applied to a wide range of systems. Finally, concluding remarks are made in Section 7.

2. Further related work

2.1. Related system models

2.1.1. Gradings

Gradings [43,35,44] form the most classical application of the overflow loss system model. It has been known for almost a century that arranging servers in a grading into a layered structure can increase the throughput of the grading. In particular, it was suggested in [19] that the best grading is the one in which there is a smooth progression "from individuals to commons", i.e. from servers with smaller skill sets to those with larger skill sets, if service time distributions are assumed to be identical for all request–server combinations. Here, we demonstrate this effect on the throughput of a more general overflow loss system model in which servers with the same skill set are combined into a *server group*.

The classic grading model [43,35,44,3] allows for arbitrary overflow policies exhibiting both inter-layer and intra-layer overflow with no spatial considerations, and is the closest model to that we consider, but permits only one server per group. Despite this restriction, accurate blocking probability evaluation remains an open problem for gradings when mutual overflow and unbalanced traffic are both present [35]. Although the exact blocking probability of an overflow loss system can be obtained in principle, by solving the underlying set of steady-state equations [25], such an approach is not scalable due to the curse of dimensionality: the number of dimensions of the state space is equal to the number of server groups in the system.

2.1.2. Cellular networks

The cellular network model considered in [36] is close to that considered in this paper, featuring both inter-layer and intra-layer overflow. However, their model is motivated by mobile cellular networks and contains strong spatial considerations, whereas in this paper we allow arbitrarily predetermined overflow policies. We also ignore the concept of handover, which is unique to wireless and cellular networks. An application of the IESA framework to cellular networks is available in [7], but this does not consider inter-layer overflow. Another model, considered in [30,31], is also similar to the one considered here, but does not allow intra-layer overflow.

2.1.3. Call centers

Avramidis et al. [45] consider a call center model with delay. As a further level of approximation, the pooling of queued requests in a common buffer is replaced with a wait-at-last-choice policy in which each server group has is own buffer for queued requests. The algorithm is used as a tool to facilitate optimization of call center staffing by reducing the amount of simulation required, and a separate analysis of its accuracy as a stand-alone performance evaluation tool is not provided.

Call centers have also been modeled in the literature using blocking models without waiting [46,40,47]. It is argued that results obtained with a loss model can be a good proxy for models with waiting [41,40]. The call center model in [46] is close to ours but does not include the concept of layering.

2.2. FPA

FPA is based on the decoupling of a given system into independent full-accessibility subsystems, for example M/M/k/k [46], in which each request offered to the subsystem may attempt every server in the subsystem. In this way, the computational time and memory requirement is greatly reduced compared to direct analysis of the entire system. Note that adoption of the Poisson assumption is equivalent to introducing an exponentially distributed delay of arbitrary mean for each overflow of a request [48, p. 157].

Such a direct decomposition approach, when applied to systems with mutual overflow, inherently gives rise to a set of interdependent non-linear equations with one or more fixed points. Higher moments (such as the variance and skewness) of overflow traffic can also be considered for obtaining more accurate approximations [49,30,37]; we shall use the term FPAm to denote FPA with moment matching.

Unfortunately, while FPAm is effective in modeling inter-layer overflow, it cannot capture the mutual dependency effects created by intra-layer overflow [32], resulting in large approximation errors in many cases [32,33]. It is demonstrated in [32] that for most systems with mutual overflow, the errors caused by the independence assumption in FPA dominate those caused by the Poisson assumption, rendering FPAm inadequate for such systems. For a thorough discussion of non-Poisson and dependence effects in overflow loss systems, see [50].

While methods of countering the effect of FPA's independence assumption appear in [11,12,51,13,14], in which correction factors are incorporated into FPA, such methods are restricted to the case of *full accessibility*, meaning that each request may attempt *all* of the servers in the system. An approximate method in [52,15] allows for both mutual overflow and limited accessibility, but only for a specific overflow policy (requests may only attempt the closest two server groups in a ring).

Finally, although FPA is guaranteed to result in at least one fixed point [53], no guarantee of convergence or uniqueness is known for FPA. Koole and Talim [46] prove convergence of FPA to a fixed point for a special case with two server groups, but do not prove uniqueness of the solution.

2.3. IESA

The IESA framework was established in [33,34,7] as a more accurate, robust and computationally efficient approximation approach to blocking probability evaluation in overflow loss systems. The framework is based on developing a *surrogate model* with a similar blocking probability to the true model and then, in a similar manner to FPA, decoupling the surrogate model into independent subsystems with Poisson input in order to deal with the curse of dimensionality.

The IESA model is based on the application of an information exchange mechanism to capture the overflow traffic dependence and hence reduce the errors caused by decoupling the surrogate model into independent subsystems with Poisson input.

In IESA, each request holds a *congestion estimate* of the number of busy server groups in its current layer, based on congestion information received by a request as it overflows from one server group to the next. This congestion information is used to identify requests with a high probability of being overflowed from the current system layer. Such requests may be *preemptively promoted* to the next system layer without attempting the remaining server groups in the current layer, or blocked if there are no more layers.

A request seeking service exchanges its congestion estimate with a request in service if and only if the request in service has a higher congestion estimate. In other words, information exchange can only increase but not decrease the congestion estimate of a request seeking service. IESA thus replaces the non-hierarchical traffic structure *within* each layer of the original overflow loss system with a hierarchical traffic structure based on this congestion estimate. Due to this hierarchical traffic structure, IESA yields a solution with guaranteed convergence and uniqueness within a finite number of iterations, whereas FPA requires a fixed-point solution. For the sake of clarity, we will use the term *level* to describe the IESA sub-hierarchy within each layer of the IESA model.

The concept of IESA is depicted in Fig. 1, showing the two main features of the IESA model: a similar but slightly higher blocking probability than the true model, and a reduction in error when decoupling is applied to the surrogate model. These two features combined result in a more accurate and robust blocking probability approximation for the true model than FPA applied directly to the true model. Intuition supporting these two features is presented at the end of this subsection.

Two IESA approximations are presented in [33] for overflow loss systems, denoted as IESA1 and IESA2. IESA1 is numerically equivalent to an earlier approximation, the Overflow Priority Classification Approximation (OPCA) [32], but replaces the preemptive priority mechanism of OPCA with an equivalent information exchange mechanism. IESA2, which uses a new surrogate model, is generally more accurate and robust than IESA1 [33]. As an extension of IESA1, IESA2 is equivalent to IESA1 in the case where each request has full access to each server in the system [54]. We shall use ''IESA'' from now on to refer specifically to IESA2 as appropriate.

More details of the IESA model are provided in Section 5.3.

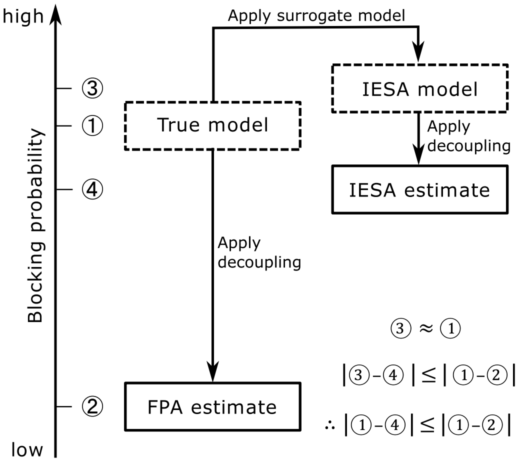

Fig. 1. Graphical depiction of the IESA framework.

2.3.1. *Intuition supporting Fig. 1*

Overflowing requests in the IESA model are preemptively promoted to the next system layer with a certain probability dependent on the system congestion estimate for the current layer, as provided by the information exchange mechanism (to be described in detail in Section 5.3). These preempted requests thus do not attempt all accessible server groups in the current system layer. As there is a non-zero probability that one of the skipped-over server groups could have served the preempted request, the blocking probability of the IESA model is slightly higher than that of the true model, i.e. ③ $\geq$ ①. On the other hand, the preempted requests are carefully chosen so that this probability is relatively low; i.e. preempted requests have a high probability of being overflowed from the layer anyway if allowed to attempt the remaining server groups in that layer. Therefore, we argue that the inequality is generally quite tight and ③ $\approx$ ①. Equality is achieved when the true model itself is purely hierarchical and there is no intra-layer overflow.

Furthermore, preemptive promotion of requests in IESA increases the proportion of the total traffic offered to a server group formed by fresh requests, whereas the proportion formed by overflowed requests is decreased (a proof of this was provided in [32, Cor. 1] for a special case). This reapportionment of the traffic offered to a server is effective in combating the errors caused by the Poisson and independence assumptions, as fresh requests to each layer have the least non-Poisson and dependency effects. In this sense, the gap between the IESA estimate and the exact blocking probability of the IESA model is narrower than that between the FPA estimate and the exact blocking probability of the true model, i.e. |③–④| $\leq$ |①–②|. In addition, the gap ③–④ is somewhat offset by the positive difference of ③ over ①. Therefore, IESA produces results closer to the real blocking probabilities than those by direct application of FPA, i.e. |①–④| $\leq$ |①–②|.

Proven theoretical bounds for IESA have been shown for a special case of overflow loss systems [32,48] in which each request may attempt all servers in a system in fully random order. In particular, IESA1 was shown for this case to always lie between FPA and the true blocking probability. IESA in this paper, equivalent to IESA2 in [33], can be shown to be equivalent to IESA1 in this special case [54]. In addition, in the case of critical loading where the total offered load in Erlangs is equivalent to the total number of servers, the ratio $B^{\text{exact}}/B^{\text{IESA}}$ between IESA and the exact blocking probability is bounded above by $\sqrt{2}$, whereas $B^{\text{exact}}/B^{\text{FPA}}$ tends to infinity as the system size increases [48].

3. Loss system model

Let L denote the number of *layers* in the overflow loss system and $\mathbf{L} = \{1, 2, \ldots, L\}$ denote the set of layers. Each layer $\ell \in \mathbf{L}$ contains a set $\mathbf{G}_\ell = \{(\ell, 1), (\ell, 2), \ldots, (\ell, G_\ell)\}$ of *server groups*. Each server group (ℓ, g) consists of $N_{\ell,g}$ servers, thus forming an M/M/$N_{\ell,g}/N_{\ell,g}$ queue. Let $\mathbf{M} = \{1, 2, \ldots, M\}$ denote the set of *request types*. Type-m requests, $m \in \mathbf{M}$, arrive to the system according to a Poisson process with rate λ_m.

Typical values of M range from 3 or 4 for an emergency care network [55,16], to several dozens for a large call center [45], to several hundreds for video-on-demand networks [56,10]. In this paper, we generally use values of $\sum_\ell G_\ell$ and M of around 100 and 500, respectively, consistent with Wong et al. [33].

Let $\left(\ell, \gamma_{m,\ell,n}\right) \in \mathbf{G}_\ell$ denote the chosen server group for type-m requests in layer ℓ, having overflowed n times so far in layer ℓ. Let $k_{m,\ell}$ denote the number of accessible server groups for type-m requests in layer ℓ, and $\Gamma_{m,\ell} = \left(\left(\ell, \gamma_{m,\ell,0}\right), \ldots, \left(\ell, \gamma_{m,\ell,k_{\ell,s}-1}\right)\right)$ denote the sequence of these server groups. Finally, let $\Gamma_m = \Gamma_{m,1} \oplus \Gamma_{m,2} \oplus \cdots \oplus \Gamma_{m,L}$ denote the entire sequence of server groups accessible to type-m requests, where $\oplus$ denotes concatenation. A type-m request will attempt each server group in Γ_m in order until an attempted server group has at least one free server, upon which the request is then served by that server. If all server groups in Γ_m are fully occupied, the request is *blocked and cleared*. The probability of such an event, known as the *blocking probability*, is an important performance measure of overflow loss systems.

Table 1
Table of notations.

Symbol	Definition
L	Number of layers in the system
$\mathbf{L}$	Set of layers in the system
G_ℓ	Number of server groups in layer ℓ
$\mathbf{G}_\ell$	Set of server groups in layer ℓ
$N_{\ell,g}$	Number of servers in server group (ℓ, g)
M	Number of request types
$\mathbf{M}$	Set of request types
λ_m	Arrival rate of type-m requests
$(\ell, \gamma_{m,\ell,n})$	Server group in layer ℓ receiving type-m requests which have overflowed n times in layer ℓ
$k_{m,\ell}$	Number of server groups in layer ℓ accessible to type-m requests
$\Gamma_{m,\ell}$	Set of server groups in layer ℓ accessible to type-m requests
Γ_m	Set of all server groups accessible to type-m requests
$A_{m,\ell}$	Offered load of type-m requests to layer ℓ, in Erlangs
$A'_{m,\ell}$	Variance of offered load of type-m requests to layer ℓ
$W_{m,\ell}$	Mean overflow of type-m requests from layer ℓ, in Erlangs
$W'_{m,\ell}$	Overflow variance of type-m requests from layer ℓ

All requests to the system are assumed to have an exponential service time distribution with unit mean. Numerical experiments in Section 6.10 suggest that the effect of assuming an exponential service time distribution is small.

Let $A_{m,\ell}$ denote the offered load in Erlangs composed of type-m requests to layer ℓ, and $W_{m,\ell}$ denote the mean overflow traffic of the type-m requests from layer ℓ. Thus

$$A_{m,\ell} = \begin{cases} \lambda_m, & \ell = 1 \\ W_{m,\ell-1}, & \ell = 2, 3, \ldots, L. \end{cases}$$

The blocking probability of type-m requests is $B_m = W_{m,L}/\lambda_m$, and the overall system blocking probability is

$$B = \frac{\sum\limits_{m \in \mathbf{M}} W_{m,L}}{\sum\limits_{m \in \mathbf{M}} A_{m,1}}.$$

The challenge of approximating B_m and B is thus reduced to that of approximating $W_{m,\ell}$ from $A_{m,\ell}$ for each layer $\ell \in \mathbf{L}$.

If moment-matching techniques are used, then more notations are required. The corresponding variances of the offered and overflow traffic for type-m requests in layer ℓ are denoted $A'_{m,\ell}$ and $W'_{m,\ell}$, respectively, where

$$A'_{m,\ell} = \begin{cases} \lambda_m, & \ell = 1 \\ W'_{m,\ell-1}, & \ell = 2, 3, \ldots, L. \end{cases}$$

A summary of the notations described in this section is provided in Table 1.

3.1. Example

An example loss system is shown in Fig. 2. In this example, there are $L = 3$ layers, with $G_1 = 3, G_2 = 4$, and $G_3 = 1$ server groups in layers 1, 2, and 3, respectively. The number of request types is $M = 4$. The number $N_{\ell,g}$ of servers in each server group (ℓ, g) and the arrival rate λ_m of traffic for each request type m is given in Fig. 2. The overflow policy is as follows: $\Gamma_{1,1} = ((1, 1))$, $\Gamma_{1,2} = ((2, 1), (2, 2))$, and $\Gamma_{1,3} = ((3, 1))$; thus $\Gamma_1 = ((1, 1), (2, 1), (2, 2), (3, 1))$. Similarly,

$$\Gamma_2 = ((1, 2), (2, 2), (2, 3), (3, 1))$$
$$\Gamma_3 = ((1, 3), (2, 3), (2, 4), (3, 1))$$
$$\Gamma_4 = ((2, 4), (2, 1), (3, 1)).$$

From this we deduce $k_{1,1} = 1, k_{1,2} = 2, k_{1,3} = 1, k_{2,1} = 1$, etc. Note that $k_{4,1} = 0$ and $\Gamma_{4,1} = ()$ (a zero-length sequence). In general, if $k_{m,\ell} = 0$ for any layer $\ell \in \mathbf{L}$, or equivalently if $\Gamma_{m,\ell} = ()$, then type-m requests in layer ℓ simply bypass the layer without seeking service in that layer, in which case $W_{m,\ell} = A_{m,\ell}$ and $W'_{m,\ell} = A'_{m,\ell}$.

4. Motivations

This section explains the motivations behind our chosen overflow loss system model, which facilitates both intra-layer mutual overflow and inter-layer hierarchical overflow.

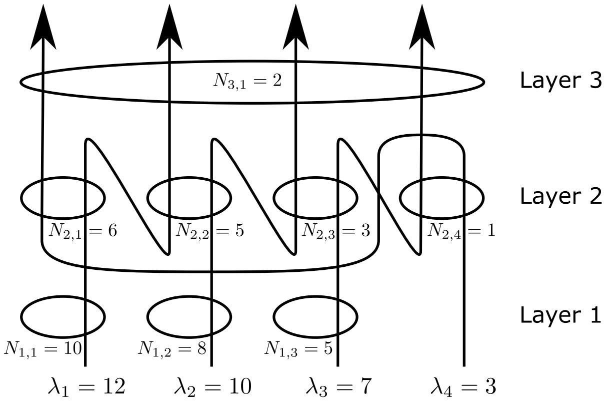

Fig. 2. Example of a three-layer overflow loss system.

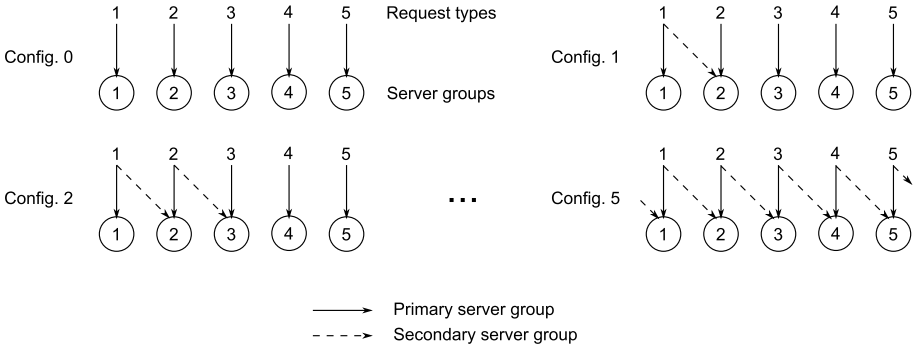

Fig. 3. Graphical depiction of Configurations 0–5 in Section 4.1.

Table 2
Blocking probabilities for Section 4.1.

Config.	Blocking prob.	95% C.I.	Ratio to previous
0	0.018369	$\pm 2.93 \times 10^{-5}$	–
1	0.015329	$\pm 3.66 \times 10^{-5}$	1.198
2	0.011920	$\pm 2.47 \times 10^{-5}$	1.286
3	0.008442	$\pm 4.36 \times 10^{-5}$	1.412
4	0.004995	$\pm 1.59 \times 10^{-5}$	1.690
5	0.001054	$\pm 4.05 \times 10^{-6}$	**4.740**

4.1. Benefits of mutual overflow

Our loss system model permits requests to attempt multiple server groups in the same system layer in an arbitrary order. This allows for the presence of mutual overflow within each system layer. To demonstrate the benefit of mutual overflow, consider an overflow loss system with $M = 5$ request types and $L = 1$ layer containing $G_1 = 5$ server groups with $N_{1,g} = 10$ servers each. The arrival rate for each request type is $\lambda_m = 5$ for all $m \in \mathbf{M}$. The overflow policies are as follows, with each subsequent configuration increasing $\sum_m k_{m,1}$ by one:

- Configuration 0: $\Gamma_m = ((1, m))$ for $m = 1, 2, 3, 4, 5$.
- Configuration i, $i = 1, 2, 3, 4, 5$: $\Gamma_m = ((1, m), (1, (m \bmod 5) + 1))$ for $m = 1, \ldots, i$; $\Gamma_m = ((1, m))$ for $m = i + 1, \ldots, 5$.

This is depicted graphically in Fig. 3. Configuration 5 thus "completes the chain" [20] and introduces mutual overflow into the system. The overall blocking probability of each scenario, as evaluated via simulation, is shown in Table 2, along with the 95% confidence interval as obtained using Student's t-distribution. The results demonstrate the benefits of allowing mutual overflow, as Configuration 5 has a considerably lower blocking probability than any other configuration. Note that while each configuration in Table 2 exhibits less blocking than the previous one, the improvement is limited compared to Configuration 5 over Configuration 4, in which increasing $\sum_m k_{m,1}$ by one suddenly results in over four times improvement.

Table 3
Blocking probabilities for Section 4.2.

$\hat{k}$	One layer		Two layers		Relative difference
	Mean	st. dev.	Mean	st. dev.	
20	0.009970	3.35×10^{-5}	0.009429	3.88×10^{-5}	−5.42%
30	0.007810	3.02×10^{-5}	0.007519	3.04×10^{-5}	−3.72%
40	0.007024	2.31×10^{-5}	0.006879	2.68×10^{-5}	−2.07%
50	0.006643	3.03×10^{-5}	0.006566	1.69×10^{-5}	−1.16%
60	0.006439	2.42×10^{-5}	0.006378	2.09×10^{-5}	−0.96%
70	0.006309	2.06×10^{-5}	0.006290	2.23×10^{-5}	−0.30%
80	0.006230	2.36×10^{-5}	0.006209	2.56×10^{-5}	−0.35%

4.2. Benefits of layering

To demonstrate the benefits of separating an overflow loss system into layers, we consider two overflow loss systems. The first system consists of a single layer ($L = 1$) of $G_1 = 100$ server groups, and the second consists of two layers ($L = 2$) with $G_1 = G_2 = 50$ server groups in each layer. All server groups in both system contain ten servers, so that each system contains a total of 1000 servers. There are $M = 500$ request types in each system, each with an arrival rate of $\lambda_m = 1.92$, so that each system receives a total offered load of 960 Erlangs or 96% loading. In the first system, each type-m request is served by $k_{m,1} = \hat{k}$ server groups. In the second system, each type-m request is served by $k_{m,1} = \hat{k}/2 + 5$ server groups in layer 1 and $k_{m,2} = \hat{k}/2 - 5$ in layer 2 (we consider even values of $\hat{k}$ only).

For each value of $\hat{k}$ in $\{20, 30, \dots, 80\}$, twenty random routing configurations are generated for both the one-layer and two-layer system, and the overall blocking probability of each configuration evaluated via simulation. The results, shown in Table 3, demonstrate consistently better performance of the two-layer system over the one-layer system. This is consistent with earlier results for gradings [19] where it is found that by creating a progression from more specialized to more generic server groups, in terms of the number of request types served, the blocking probability of an overflow loss system can be reduced.

5. Approximation

5.1. FPA

FPA makes two major simplifying assumptions: that the traffic offered to each server group in the system is Poisson and independent of the traffic offered to the other server groups. However, since the traffic offered to each server group still depends on the offered traffic to and blocking probability of the other server groups, fixed-point iteration [53] is required to find the traffic offered to each group.

Consider layer ℓ on its own. Define:

- $a_{m,\ell,n}$—Offered traffic composed of type-m requests, having overflowed n times so far in layer ℓ, for all $m \in \mathbf{M}$, $0 \le n < k_{m,\ell}$. These requests are always offered to server group $\left(\ell, \gamma_{m,\ell,n}\right)$.
- $a_{\ell,g}$—Total offered traffic to group (ℓ, g), for all $g \in \mathbf{G}_\ell$.
- $b_{\ell,g}$—Congestion probability of group (ℓ, g), for all $g \in \mathbf{G}_\ell$, namely the probability that all servers in group (ℓ, g) are occupied.

Summing over all eligible $m \in \mathbf{M}$, we obtain

$$a_{\ell,g} = \sum_{m \in \mathbf{M}} \sum_{n=0}^{k_{m,\ell}-1} \mathbf{1}\left\{\gamma_{m,\ell,n} = g\right\} a_{m,\ell,n},$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function. From the Poisson assumption, the blocking probability of group (ℓ, g) is estimated via the Erlang B formula: $b_{\ell,g} = E\left(a_{\ell,g}, N_{\ell,g}\right)$. From the independence assumption,

$$a_{m,\ell,n} = \begin{cases} A_{m,\ell}, & n = 0; \\ a_{m,\ell,n-1} b_{\ell,\gamma_{m,\ell,n-1}}, & n = 1, 2, \dots, k_{m,\ell} - 1. \end{cases}$$

The above equations form a fixed-point system which may be solved iteratively. The overflow traffic of type-m requests from layer ℓ is

$$W_{m,\ell} = A_{m,\ell} \prod_{n=0}^{k_{m,\ell}-1} b_{\ell,\gamma_{m,\ell,n}} = a_{m,\ell,k_{m,\ell}-1} b_{\ell,\omega} \tag{1}$$

where $\omega = \gamma_{m,\ell,k_{m,\ell}-1}$. Eq. (1) can be interpreted simply as follows: the overflow traffic of type m requests from a layer is equal to the offered load of type-m requests to that layer, reduced by the congestion probabilities of all accessible server groups in that layer, namely those in $\Gamma_{m,\ell}$.

In this paper, we set a stopping criterion for FPA as follows: Let $b_{\ell,g}^{(n)}$ represent the nth-iteration estimate of $b_{\ell,g}$. If $|b_{\ell,g}^{(n)} - b_{\ell,g}^{(n-1)}| < 10^{-8}$ for *all* server groups (ℓ, g) in the system, then FPA is concluded and the nth-iteration estimates are used as the final estimates.

5.2. FPAm

We eliminate the Poisson assumption of FPA, but assume that overflow traffic can be adequately characterized by its mean and variance only. Consider layer ℓ on its own. Define $a_{m,\ell,n}$, $a_{\ell,g}$, and $b_{\ell,g}$ as for FPA, and $a'_{m,\ell,n}$ and $a'_{\ell,g}$ as the corresponding variances. Summing over all eligible $m \in \mathbf{M}$, we obtain

$$a_{\ell,g} = \sum_{m\in\mathbf{M}} \sum_{n=0}^{k_{m,\ell}-1} \mathbf{1}\left\{\gamma_{m,\ell,n} = g\right\} a_{m,\ell,n}$$

$$a'_{\ell,g} = \sum_{m\in\mathbf{M}} \sum_{n=0}^{k_{m,\ell}-1} \mathbf{1}\left\{\gamma_{m,\ell,n} = g\right\} a'_{m,\ell,n}.$$

The blocking probability of group (ℓ, g) is estimated via Hayward's approximation:

$$b = E\left(a_{\ell,g}, a'_{\ell,g}, N_{\ell,g}\right) = E\left(\frac{a_{\ell,g}}{z_{\ell,g}}, \frac{N_{\ell,g}}{z_{\ell,g}}\right)$$

where $z_{\ell,g} = a'_{\ell,g}/a_{\ell,g}$. For details on extending the Erlang B formula to non-integer number of servers, see [57]. From the independence assumption,

$$a_{m,\ell,n} = \begin{cases} A_{m,\ell}, & n = 0; \\ a_{m,\ell,n-1} b_{\ell,\gamma_{m,\ell,n-1}}, & n = 1, 2, \ldots, k_{m,\ell} - 1. \end{cases}$$

$$a'_{m,\ell,n} = \begin{cases} A'_{m,\ell}, & n = 0; \\ \mathcal{M}\left(a_{m,\ell,n-1}, a'_{m,\ell,n-1}, b_{\ell,\gamma_{m,\ell,n-1}}\right), & n = 1, 2, \ldots, k_{m,\ell} - 1, \end{cases}$$

where $\mathcal{M}$ depends on the chosen moment-matching method. In this paper, we choose a moment-matching method proposed by Huang et al. [31], which we describe in detail in Section 5.7.

The above equations form a fixed-point system which may be solved iteratively. The overflow traffic of type-m requests from layer ℓ is

$$W_{m,\ell} = A_{m,\ell} \prod_{n=0}^{k_{m,\ell}-1} b_{\ell,\gamma_{m,\ell,n}} = a_{m,\ell,k_{m,\ell}-1} b_{\ell,\omega}$$

with corresponding variance

$$W'_{m,\ell} = \mathcal{M}\left(a_{m,\ell,k_{m,\ell}-1}, a'_{m,\ell,k_{m,\ell}-1}, b_{\ell,\omega}\right),$$

where $\omega = \gamma_{m,\ell,k_{m,\ell}-1}$.

5.3. IESA—Basic description

IESA involves applying the same methodology as FPA, namely decoupling of a system into full-accessibility subsystems with Poisson input, but while FPA applies decoupling to the true model as defined in Section 3, IESA applies decoupling to a *surrogate model* instead. This surrogate model, which we call the IESA model, is designed so that the non-hierarchical dependencies inherent in the true model are captured within the hierarchy of the IESA model. As a result, IESA exhibits much smaller errors than FPA in many cases.

In the IESA model, each request has three parameters: m, its type, Δ, its set of previously attempted server groups in the current layer, and Ω, its *congestion estimate* (a scalar) of the number of fully occupied server groups in the current layer. A high Ω value means high system congestion and hence high system interdependence, meaning that if a request finds that a server group to be full, that request is likely to find that other server groups are also full. Classification of requests within each system layer by their Ω value creates a sub-hierarchy *within* each system layer where the jth level of the sub-hierarchy includes incoming requests for which $\Omega \leq j$. In other words, Ω forms the mechanism by which IESA captures mutual dependencies between server groups in a hierarchical manner.

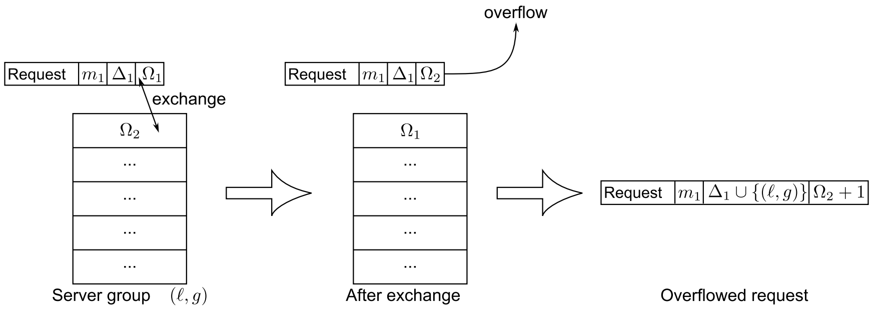
overflow
Request m_1 Δ_1 Ω_1
exchange
Ω_2
...
Server group (ℓ, g)
Request m_1 Δ_1 Ω_2
Ω_1
After exchange
Request m_1 $\Delta_1 \cup \{(\ell, g)\}$ $\Omega_2 + 1$
Overflowed request

Fig. 4. Graphical depiction of the information exchange mechanism in the IESA model. In this figure, Ω_2 represents the highest Ω value of all requests in service at server group (ℓ, g), with $\Omega_1 < \Omega_2$.

Consider layer ℓ on its own. All incoming requests to layer ℓ start with $\Delta = \emptyset$ and $\Omega = 0$. When an incoming $(m_1, \Delta_1, \Omega_1)$-request encounters a fully occupied server group (ℓ, g), it compares itself to the most senior (highest Ω value) request in service, which we denote as an $(m_2, \Delta_2, \Omega_2)$-request. Ties are broken arbitrarily. If $\Omega_1 \geq \Omega_2$, then no information exchange occurs and the incoming request overflows as an $(m_1, \Delta_1 \cup \{(\ell, g)\}, \Omega_1 + 1)$-request. On the other hand, if $\Omega_1 < \Omega_2$, then the incoming request exchanges its Ω value with the request in service and overflows as an $(m_1, \Delta_1 \cup \{(\ell, g)\}, \Omega_2 + 1)$-request, while the request in service becomes an $(m_2, \Delta_2, \Omega_1)$-request. In this way, $|\Delta| \leq \Omega$ for all incoming calls. A graphical depiction of this information exchange mechanism is given in Fig. 4.

For overflowing (m, Δ, j)-requests in layer ℓ, $|\Delta| = n$, there is a certain probability $P_{m,\ell,n,j}$ that the $k_{m,\ell} - n$ unvisited server groups in $\Gamma_{m,\ell} \setminus \Delta$ are all fully occupied. We estimate $P_{m,\ell,n,j}$ by considering Erlang's Ideal grading [58] with $G_\ell - n$ individual servers, of which $k_{m,\ell} - n$ servers may be attempted at random and $j - n$ servers are currently occupied. Thus

$$P_{m,\ell,n,j} = \begin{cases} \dfrac{\binom{j-n}{k_{m,\ell}-n}}{\binom{G_\ell - n}{k_{m,\ell}-n}}, & j \geq k_{m,\ell}; \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

This estimate is used by the IESA model to control overflow in the system. With probability $1 - P_{m,\ell,n,j}$, the request is offered as normal to the next server group in its overflow policy, namely $\left(\ell, \gamma_{m,\ell,n}\right)$. With probability $P_{m,\ell,n,j}$, the request immediately overflows to the next system layer without attempting any server groups in $\Gamma_{m,\ell} \setminus \Delta$. If there are no more layers, the request is blocked and cleared. Note that $P_{m,\ell,n,j} = 1$ when $n = k_{m,\ell}$ (each server group in $\Gamma_{m,\ell}$ has been visited) or when $j = G_\ell$ (all server groups in $\mathbf{G}_\ell$ are *believed* to be fully occupied). Also, as $n \leq j$ for all incoming requests, $0 \leq P_{m,\ell,n,j} \leq 1$, confirming that our definition of $P_{m,\ell,n,j}$ is a valid one.

In summary, IESA transforms *each* layer of the true model from a non-hierarchical traffic dependency structure to a purely hierarchical traffic dependency structure based on Ω, resulting in provable convergence of IESA to a unique solution, which FPA does not provide. The hierarchical dependency structure created by the IESA model (i.e. the IESA2 model in [33]) is superior to that created by the IESA1 model, where the *identity* of a request is also exchanged in addition to the congestion estimate, creating a surrogate model that is further from reality than the IESA2 model.

5.4. IESA—detailed description

Consider layer ℓ on its own and define $e_{m,\ell,n,j}$, $\tilde{e}_{m,\ell,n,j}$, $x_{m,\ell,n,j}$, $w_{m,\ell,n,j}$, $a_{\ell,g,n,j}$, $\tilde{a}_{\ell,g,n,j}$, $a_{\ell,g,j}$, and $b_{\ell,g,j}$ as in Table 4. For recursion purposes, all values above are assumed to be zero outside of the allowed indices. By definition:

$$\begin{aligned} w_{m,\ell,n,j} &= x_{m,\ell,n,j} P_{m,\ell,n,j} \\ e_{m,\ell,n,j} &= x_{m,\ell,n,j}\left(1 - P_{m,\ell,n,j}\right), \quad n > 0 \\ \tilde{e}_{m,\ell,n,j} &= \sum_{i=n}^{j} e_{m,\ell,n,i} \\ \tilde{a}_{\ell,g,n,j} &= \sum_{i=n}^{j} a_{\ell,g,n,i}. \end{aligned}$$

As only fresh calls can have $|\Delta| = 0$, we obtain

$$e_{m,\ell,0,j} = \begin{cases} A_{m,\ell}, & j = 0 \\ 0, & j = 1, 2, \ldots, G_\ell - 1. \end{cases}$$

Table 4
Table of notations for IESA.

Symbol	Definition	Allowed indices				
		m	ℓ	n	j	g
$e_{m,\ell,n,j}$	Total offered traffic composed of (m,Δ,j)-requests, $\lvert\Delta\rvert = n$	$m \in \mathbf{M}$	$0\ldots L-1$	$0\ldots k_{m,\ell}-1$	$n\ldots G_\ell-1$	–
$\tilde{e}_{m,\ell,n,j}$	Total offered traffic composed of (m,Δ,Ω)-requests, $\lvert\Delta\rvert = n$, $n \le \Omega \le j$	$m \in \mathbf{M}$	$0\ldots L-1$	$0\ldots k_{m,\ell}-1$	$n\ldots G_\ell-1$	–
$x_{m,\ell,n,j}$	Total overflow traffic from server group $\left(\ell,\gamma_{m,\ell,n-1}\right)$ composed of (m,Δ,j)-requests, $\lvert\Delta\rvert = n$.	$m \in \mathbf{M}$	$0\ldots L-1$	$1\ldots k_{m,\ell}$	$n\ldots G_\ell$	–
$w_{m,\ell,n,j}$	Portion of $x_{m,\ell,n,j}$ which is preempted and overflows immediately to layer $\ell+1$ (or blocked and cleared if $\ell = L$)	$m \in \mathbf{M}$	$0\ldots L-1$	$1\ldots k_{m,\ell}$	$n\ldots G_\ell$	–
$a_{\ell,g,n,j}$	Total offered traffic to server group (ℓ,g) composed of (m,Δ,j)-requests, $m \in \mathbf{M}$ and $\lvert\Delta\rvert = n$	–	$0\ldots L-1$	$0..k^\star_{\ell,g}-1$	$n\ldots G_\ell-1$	$g \in \mathbf{G}_\ell$
$\tilde{a}_{\ell,g,n,j}$	Total offered traffic to server group (ℓ,g) composed of (m,Δ,Ω)-requests, $m \in \mathbf{M}$, $\lvert\Delta\rvert = n$, and $n < \Omega < j$	–	$0\ldots L-1$	$0..k^\star_{\ell,g}-1$	$n\ldots G_\ell-1$	$g \in \mathbf{G}_\ell$
$a_{\ell,g,j}$	Total offered traffic to server group (ℓ,g) composed of (m,Δ,Ω)-requests, $m \in \mathbf{M}$ and $0 \le \lvert\Delta\rvert \le \Omega \le j$	–	$0\ldots L-1$	–	$0\ldots G_\ell-1$	$g \in \mathbf{G}_\ell$
$b_{\ell,g,j}$	Congestion probability of server group (ℓ,g) at level j of the IESA hierarchy in layer ℓ; in other words the probability that all servers in (ℓ,g) are occupied by calls with $\Omega \le j$	–	$0\ldots L-1$	–	$0\ldots G_\ell-1$	$g \in \mathbf{G}_\ell$

$k^\star_{\ell,g} = \max_{m:(\ell,g)\in\Gamma_{m,\ell}}$ denotes the largest value of $k_{m,\ell}$ for any request type with access to server group (ℓ,g).

Summing over all eligible $m \in \mathbf{M}$,

$$a_{\ell,g,n,j} = \sum_{m\in\mathbf{M}} \mathbf{1}\left\{\gamma_{m,\ell,n} = g\right\} e_{m,\ell,n,j}.$$

Let $k^\star_{\ell,g} = \max_{m:(\ell,g)\in\Gamma_{m,\ell}} k_{m,\ell}$ denote the largest value of $k_{m,\ell}$ for all request types m with access to server group (ℓ,g). Then

$$a_{\ell,g,j} = \sum_{n=0}^{\min\left(j,k^\star_{\ell,g}-1\right)} \tilde{a}_{\ell,g,n,j}.$$

From the Poisson assumption, the blocking probability of group (ℓ,g) is estimated via the Erlang B formula: $b_{\ell,g,j} = E\left(a_{\ell,g,j}, N_{\ell,g,j}\right)$.

Consider the server group $(\ell,g) = \left(\ell,\gamma_{m,\ell,n-1}\right)$ for some $m \in \mathbf{M}$ and $1 \le n \le k_{m,\ell}$. In our IESA model, there are two ways for a request to overflow from (ℓ,g) with a congestion estimate of j:

1. A request with congestion estimate $j-1$ arriving at (ℓ,g) finds with probability $b_{\ell,g,j-1}$ 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 simply overflows with congestion estimate j.
2. A request with a congestion estimate of $i \le j-2$ finds with probability $b_{\ell,g,j-1} - b_{\ell,g,j-2}$ that all servers are busy, with the most senior request in service having a congestion estimate of exactly $j-1$. Since the incoming request is junior (smaller congestion estimate) to this request, the congestion estimates of the two requests are exchanged. The request in service obtains a new congestion estimate of i, while the incoming request overflows with a congestion estimate of j.

Combining these two possibilities, we obtain

$$x_{m,\ell,n,j} = e_{m,\ell,n-1,j-1} b_{\ell,\gamma_{m,\ell,n-1},j-1} + \tilde{e}_{m,\ell,n-1,j-2}\left(b_{\ell,\gamma_{m,\ell,n-1},j-1} - b_{\ell,\gamma_{m,\ell,n-1},j-2}\right).$$

The above values can be obtained iteratively for $j = 0, 1, \ldots, G_{\ell-1}$. The overflow traffic of type-m requests from layer ℓ is

$$W_{m,\ell} = \sum_{n=1}^{k_{m,\ell}} \sum_{j=k_{m,\ell}}^{G_\ell} w_{m,\ell,n,j}.$$

To further explain the derivation of the IESA algorithm for our overflow loss system model, the relation between $a_{\ell,g,j}$, $b_{\ell,g,j}$, and the IESA hierarchy is illustrated in Fig. 5. The proportion of requests at each level which are immediately overflowed to the next layer depends on the Δ values of the individual requests.

5.5. IESAm

Assume that overflow traffic can be adequately characterized by its mean and variance only. Consider layer ℓ on its own. Define $e_{m,\ell,n,j}$, $\tilde{e}_{m,\ell,n,j}$, $x_{m,\ell,n,j}$, $w_{m,\ell,n,j}$, $a_{\ell,g,n,j}$, $\tilde{a}_{\ell,g,n,j}$, $a_{\ell,g,j}$, and $b_{\ell,g,j}$ as for IESA and $e'_{m,\ell,n,j}$, $\tilde{e}'_{m,\ell,n,j}$, $x'_{m,\ell,n,j}$, $w'_{m,\ell,n,j}$, $a'_{\ell,g,n,j}$,

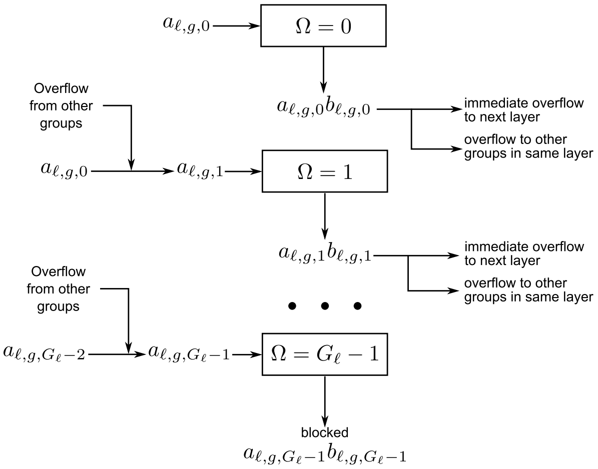
$a_{\ell,g,0}$
$\Omega = 0$
Overflow from other groups
$a_{\ell,g,0}b_{\ell,g,0}$
immediate overflow to next layer
overflow to other groups in same layer
$a_{\ell,g,0}$
$a_{\ell,g,1}$
$\Omega = 1$
$a_{\ell,g,1}b_{\ell,g,1}$
immediate overflow to next layer
overflow to other groups in same layer
Overflow from other groups
$a_{\ell,g,G_\ell-2}$
$a_{\ell,g,G_\ell-1}$
$\Omega = G_\ell - 1$
blocked
$a_{\ell,g,G_\ell-1}b_{\ell,g,G_\ell-1}$

Fig. 5. Graphical depiction of the IESA hierarchy, showing the offered load to and overflow from server group (ℓ, g) at each level Ω of the IESA hierarchy.

$\tilde{a}'_{\ell,g,n,j}$, and $a'_{\ell,g,j}$ as the corresponding variances. By definition:

$$\begin{aligned}
w_{m,\ell,n,j} &= x_{m,\ell,n,j}P_{m,\ell,n,j}\\
w'_{m,\ell,n,j} &= x'_{m,\ell,n,j}P_{m,\ell,n,j}\\
e_{m,\ell,n,j} &= x_{m,\ell,n,j}\left(1 - P_{m,\ell,n,j}\right), \quad n > 0\\
e'_{m,\ell,n,j} &= x'_{m,\ell,n,j}\left(1 - P_{m,\ell,n,j}\right), \quad n > 0\\
\tilde{e}_{m,\ell,n,j} &= \sum_{i=n}^{j} e_{m,\ell,n,i}\\
\tilde{e}'_{m,\ell,n,j} &= \sum_{i=n}^{j} e'_{m,\ell,n,i}\\
\tilde{a}_{\ell,g,n,j} &= \sum_{i=n}^{j} a_{\ell,g,n,i}\\
\tilde{a}'_{\ell,g,n,j} &= \sum_{i=n}^{j} a'_{\ell,g,n,i}.
\end{aligned}$$

As only fresh calls can have $|\Delta| = 0$, we obtain

$$\begin{aligned}
e_{m,\ell,0,j} &= \begin{cases} A_{m,\ell}, & j = 0\\ 0, & j = 1, 2, \ldots, G_\ell - 1 \end{cases}\\
e'_{m,\ell,0,j} &= \begin{cases} A'_{m,\ell}, & j = 0\\ 0, & j = 1, 2, \ldots, G_\ell - 1. \end{cases}
\end{aligned}$$

Furthermore,

$$\begin{aligned}
a_{\ell,g,n,j} &= \sum_{m\in\mathbf{M}} \mathbf{1}\left\{\gamma_{m,\ell,n} = g\right\} e_{m,\ell,n,j}\\
a'_{\ell,g,n,j} &= \sum_{m\in\mathbf{M}} \mathbf{1}\left\{\gamma_{m,\ell,n} = g\right\} e'_{m,\ell,n,j}\\
a_{\ell,g,j} &= \sum_{n=0}^{\min\left(j,k^{\star}_{\ell,g}-1\right)} \tilde{a}_{\ell,g,n,j}\\
a'_{\ell,g,j} &= \sum_{n=0}^{\min\left(j,k^{\star}_{\ell,g}-1\right)} \tilde{a}'_{\ell,g,n,j}.
\end{aligned}$$

The blocking probability of group (ℓ, g) is estimated via Hayward's approximation:

$$b_{\ell,g,j} = E\left(\frac{a_{\ell,g,j}}{z_{\ell,g,j}}, \frac{N_{\ell,g,j}}{z_{\ell,g,j}}\right)$$

where $z_{\ell,g,j} = a'_{\ell,g,j}/a_{\ell,g,j}$. Then

$$x_{m,\ell,n,j} = e_{m,\ell,n-1,j-1} b_{\ell,\gamma_{m,\ell,n-1},j-1} + \tilde{e}_{m,\ell,n-1,j-2}\left(b_{\ell,\gamma_{m,\ell,n-1},j-1} - b_{\ell,\gamma_{m,\ell,n-1},j-2}\right)$$
$$x'_{m,\ell,n,j} = \mathcal{M}\left(e_{m,\ell,n-1,j-1}, e'_{m,\ell,n-1,j-1}, b_{\ell,\gamma_{m,\ell,n-1},j-1}\right) + \mathcal{M}\left(\tilde{e}_{m,\ell,n-1,j-2}, \tilde{e}'_{m,\ell,n-1,j-2}, b_{\ell,\gamma_{m,\ell,n-1},j-1} - b_{\ell,\gamma_{m,\ell,n-1},j-2}\right)$$

where $\mathcal{M}$ depends on the chosen moment-matching method. In this paper, we choose a moment-matching method proposed by Huang et al. [31], which we describe in detail in Section 5.7.

The above values can be obtained iteratively for $j = 0, 1, \ldots, G_{\ell-1}$. The overflow traffic of type-m requests from layer ℓ is

$$W_{m,\ell} = \sum_{n=1}^{k_{m,\ell}} \sum_{j=k_{m,\ell}}^{G_\ell} w_{m,\ell,n,j}$$

with corresponding variance

$$W'_{m,\ell} = \sum_{n=1}^{k_{m,\ell}} \sum_{j=k_{m,\ell}}^{G_\ell} w'_{m,\ell,n,j}.$$

5.6. IESA$^+$ *and* IESAm$^+$

Eq. (2) relies on the implicit assumption that Ω and Δ are perfectly correlated: all server groups in Δ are accounted for in the value of Ω. Due to information exchange, this is not necessarily the case. We therefore propose a new IESA surrogate model where $P_{m,\ell,n,j}$ is replaced by $P^+_{m,\ell,n,j}$, in which Ω and Δ are assumed to be *independent*. While $P_{m,\ell,n,j}$ is based on an Erlang's Ideal Grading (EIG) in which the n visited server groups are excluded, $P^+_{m,\ell,n,j}$ is based on an EIG which *does* include the n visited groups. The EIG on which $P^+_{m,\ell,n,j}$ is based includes G_ℓ individual servers, of which $k_{m,\ell} - n$ servers may be attempted at random and Ω servers are currently occupied. Thus

$$P^+_{m,\ell,n,j} = \begin{cases} \dfrac{\binom{j}{k_{m,\ell}-n}}{\binom{G_\ell}{k_{m,\ell}-n}}, & j \geq k_{m,\ell}; \\ 0, & \text{otherwise}. \end{cases} \tag{3}$$

We shall use the term "IESA$^+$ model" to refer to the new surrogate model, and IESA$^+$ and IESAm$^+$ to refer to the resulting approximations with and without moment matching.

5.7. *Moment matching*

In this paper, we use the moment matching method of Huang et al. [31]. As we do not consider request requiring multiple service units, we present a simplified version of the original method here. The method of [31] computes the overflow mean and variance of separate traffic substreams offered to a server group by modeling the server group as a collection of imaginary M/M/n/n server groups, one for each substream of the combined offered traffic.

Consider a server group with N servers offered traffic with a mean of A and a variance of A', with $Z = A'/A$. The blocking probability of the server group is estimated using Hayward's approximation as $B = E\,(A/Z, N/Z)$. We are interested in the overflow process of a particular substream with an offered mean of a and an offered variance of a'. The overflow mean of the substream is $w = aB$. Let $z = a'/a$.

To calculate the overflow variance w', we construct an imaginary M/M/n/n server group with an offered load of $\varphi = a/z$ and n servers such that $E\,(\varphi, n) = B$. The overflow mean of this imaginary group is $\chi = \varphi B$, while the overflow variance is computed via Riordan's formula [28, Appx. I]:

$$\chi' = \chi\left[1 - \chi + \frac{\varphi}{n + 1 + \chi - \varphi}\right].$$

Finally, w' is estimated as $\chi' z$. We shall write $w' = \mathcal{M}\left(a, a', B\right)$. Note that if the substream consists of the entire offered load, i.e. $a = A$ and $a' = A'$, then the method becomes equivalent to that of Fredericks [29]. A graphical representation of our method is given in Fig. 6.

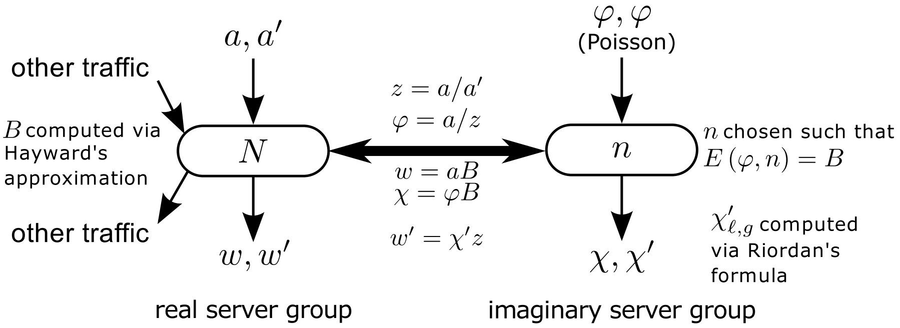

Fig. 6. Graphical depiction of the moment matching method used in this paper.

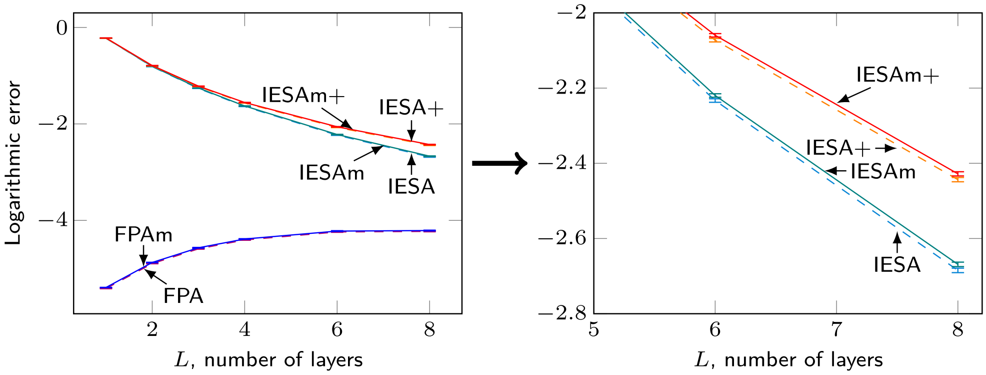

Fig. 7. Logarithmic errors for the scenarios described in Section 6.1.

6. Numerical results

This section examines the performance of FPA, FPAm, IESA, IESAm, IESA$^+$, and IESAm$^+$ as applied to the overflow loss model described in Section 3. We consider systems with two or more layers, with the skill set of servers in each layer being, on average, at least as large as those in the previous layer, consistent with the design principles of [19]. For each data point in each graph in this section, twenty (unless otherwise stated) random routing configurations are generated. For each configuration, the overall blocking probability of the system is evaluated via each approximation and compared against simulation results. In Section 6.9, individual request blocking probabilities are also considered. The simulation values for each routing configuration are obtained by conducting the following:

- A minimum of five simulation runs of 50 million request arrivals each.
- Additional simulation runs until the 95% confidence interval, as obtained using Student's t-distribution, is less than one percent of the simulation mean, or until fifteen runs have been completed.

Each data point shows the mean logarithmic error of each approximation, with error bars representing one standard deviation. Logarithmic error is defined as

$$\log_{10}\left(\frac{\text{approximation blocking probability}}{\text{simulation blocking probability}}\right).$$

6.1. Varying the number of layers

We construct an overflow loss system with 96 server groups split evenly across $L = 1, 2, 3, 4, 6$, or 8 layers, with $N_{\ell,g} = 10$ for all server groups (ℓ, g). Each request may attempt half of the server groups in each layer: $k_{m,\ell} = G_\ell/2$ for all m and ℓ. There are $M = 500$ request types, each with an arrival rate of $\lambda_m = 920/M$ Erlangs. The results are shown in Fig. 7.

All IESA approximations are demonstrated to be several orders of magnitude more accurate than FPA and FPAm, with the largest benefits when L is small. The convergence of the IESA approximations to FPA as L increases makes intuitive sense as in the extreme case of $L = 48$, $k_{m,\ell} = 1$ for all request types k and layers ℓ and thus IESA and FPA must be equal. Furthermore, IESA$^+$ produces a notably higher estimate than IESA, with the difference increasing in L (although IESA and IESA$^+$ must eventually converge when $L = 48$). IESAm$^+$ is demonstrated to be the most accurate approximation for all values of L.

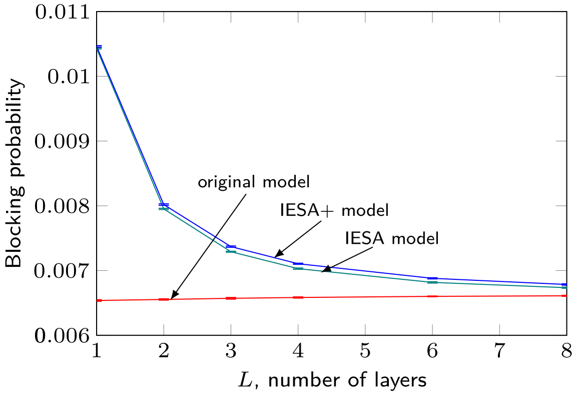

Fig. 8. Simulated blocking probabilities of the three surrogate models for the scenarios described in Section 6.1.

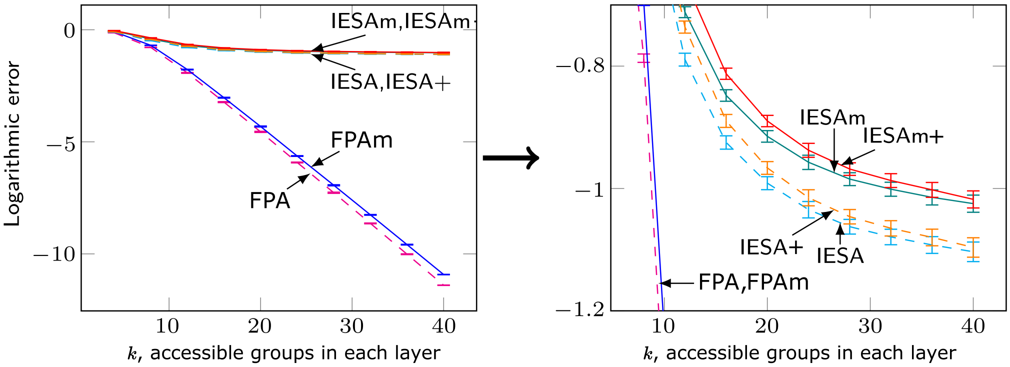

Fig. 9. Logarithmic errors for the scenarios described in Section 6.2.

Moment matching is demonstrated to have a small but consistent effect on the accuracy of FPA, IESA, and IESA$^+$. Nevertheless, the accuracy of IESAm and IESAm$^+$ is decreasing in L.

The blocking probabilities of the IESA and IESA$^+$ models, as evaluated via simulation, are shown in Fig. 8. It is demonstrated that the two surrogate models have blocking probabilities slightly higher than but similar to that of the true model, with the difference decreasing as the number of layers increases (i.e. as the number of server groups in each layer decreases).

6.2. Varying the accessibility

We consider a case with $L = 2$ layers, $G_1 = 60, G_2 = 40, N_{\ell,g} = 10$ for all server groups $(\ell, g), M = 500$, and $\lambda_m = 960/M$ for all request types m. Each request of type m may attempt $k_{m,1}$ server groups in layer 1 and $k_{m,2}$ server groups in layer 2. We set $k_{m,1} = k_{m,2} = k$ for various values of k. The results are shown in Fig. 9.

The results demonstrate a rapid deterioration of FPA and FPAm as k increases. This is because as k increases, the number of request types served by each server group also increases, in turn increasing the interdependencies between server groups in each layer. In addition, the effect of the Poisson assumption is amplified both due to the increased peakedness of overflow traffic and due to the cascading effect: any error in estimating the overflow traffic of a request type after a given number of overflows affects the offered traffic of all subsequent server group attempts.

All four IESA approximations shown demonstrate a vast improvement in accuracy over FPA and FPAm, with an improvement of roughly ten orders of magnitude for $k = 40$. Moment matching is demonstrated to provide a small additional benefit, with IESAm$^+$ the most accurate of approximation for all values of k shown.

Finally, Fig. 10 shows the mean running time for FPAm and each IESA approximation for each value of k, with the standard deviation shown as error bars. FPA is not shown as the running time of FPA is less than the resolution of the system clock on the machine on which these configurations were evaluated. It is demonstrated that IESA and IESA$^+$ have nearly identical running times, as do IESAm and IESAm$^+$. The running times of IESAm and FPAm are similar, with FPAm faster for small k and IESAm faster for large k.

The results in Fig. 10 were obtained on an IBM server with two Intel Xeon CPUs running at 2.6 GHz with 96 GB of RAM. For comparison, Markov-chain simulation consistently required at least 6000 s for each configuration, a difference of over two orders of magnitude more than FPAm and IESAm.

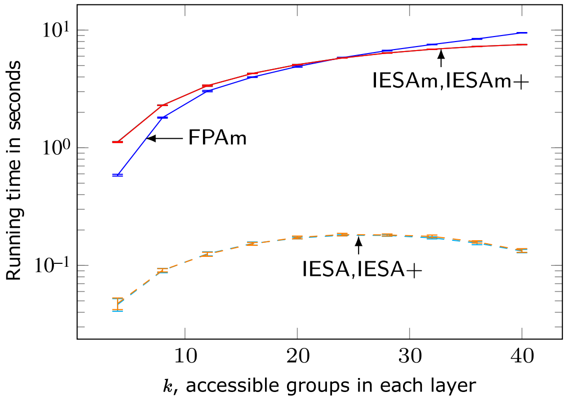

Fig. 10. Running time per configuration for the scenarios described in Section 6.2. The running time of FPA is less than the resolution of the system clock on the machine on which these configurations were evaluated.

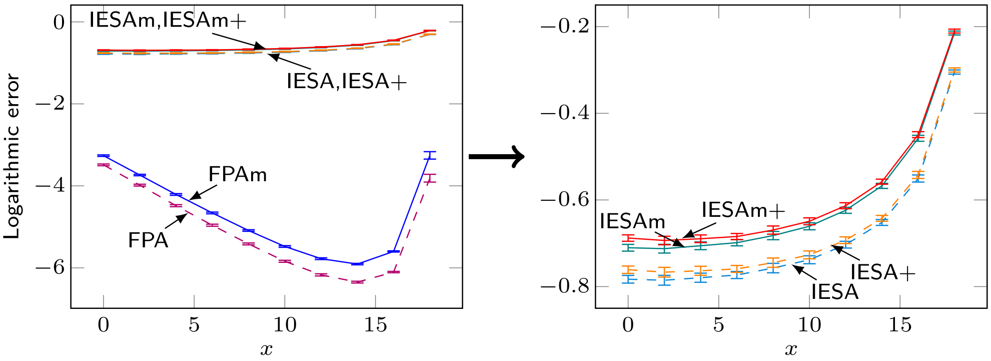

Fig. 11. Logarithmic errors for the scenarios described in Section 6.3. Each request may attempt $20-x$ server groups in layer 1 and $20+x$ groups in layer 2.

6.3. Varying the accessibility of each layer

In this subsection, we maintain $L = 2$ and $G_1 = G_2 = 50$ but varying the values of $k_{m,1}$ and $k_{m,2}$ so that $k_{m,1} = 20 - x$ and $k_{m,2} = 20 + x$ for all request types m. We also maintain $M = 500$, $\lambda_m = 960/M$ for all request types m, and $N_{\ell,g} = 10$ for all server groups (ℓ, g). The results are shown in Fig. 11.

The results demonstrate a deterioration of FPA and FPAm as x increases, except for large x where the higher blocking probability of the system becomes a factor. Moment matching is shown to have a small positive effect, slightly reducing the error of IESAm and IESAm$^+$ compared to IESA and IESA$^+$, respectively. IESAm$^+$ is demonstrated to be the most accurate approximation for all values of G_2 shown.

6.4. Varying the number of server groups in each layer

In this subsection, we consider a two-layer ($L = 2$) overflow loss system with $G_1 + G_2 = 100$ server groups in total, $G_1 \geq G_2$, and examine the effect of assigning different numbers of groups to each layer while keeping the number of groups accessible to each request in each layer the same. Each server group consists of $N_{\ell,g} = 10$ identical servers. There are $M = 500$ request types with an arrival rate of $\lambda_m = 960/M$ Erlangs each, and requests of each type m may attempt $k_{m,1} = k_{m,2} = 20$ server groups in each layer. The results are shown in Fig. 12.

The results demonstrate a sharp deterioration of FPA and FPAm as G_2 decreases, with the IESA approximations outperforming FPA by roughly seven orders of magnitude for $G_2 = 20$. This is because as G_2 decreases, the number of request types served by each server group in layer 2 increases, which in turn increases the interdependencies between server groups in the layer. Moment matching is shown to have a small positive effect, slightly reducing the error of IESAm and IESAm$^+$ compared to IESA and IESA$^+$, respectively. IESAm$^+$ is demonstrated to be the most accurate approximation for all values of G_2 shown.

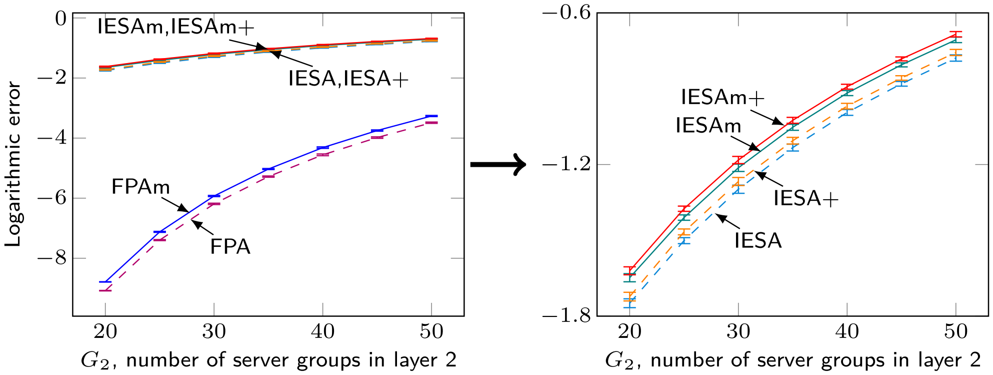

Fig. 12. Logarithmic errors for the scenarios described in Section 6.4.

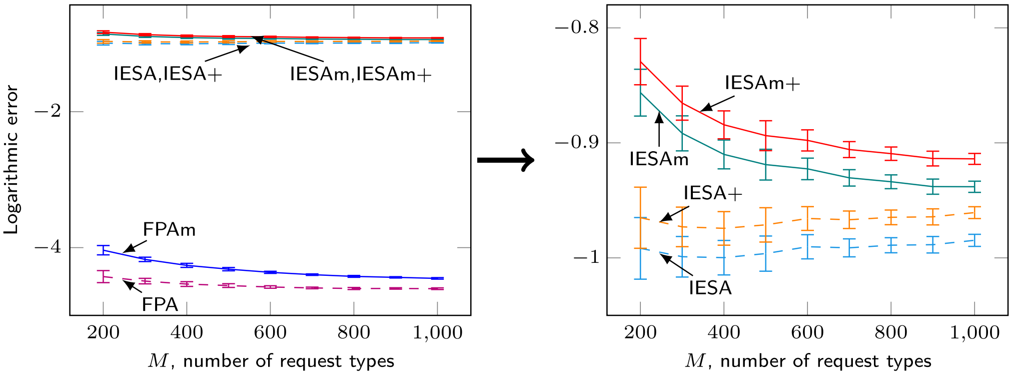

Fig. 13. Logarithmic errors for the scenarios described in Section 6.5.

6.5. *Varying the number of request types*

We consider a case with $L = 2$ layers, with $G_1 = 60$ and $G_2 = 40$. Each server group contains $N_{\ell,g} = 10$ servers. We vary the number of request types M, each with an arrival rate of $\lambda_m = 960/M$. Each request may attempt $k_{m,1} = k_{m,2} = 20$ server groups in each layer. The results are shown in Fig. 13.

The IESA approximations are consistently more accurate than FPA and FPAm by several orders of magnitude. Moment matching is shown to have a small positive effect, slightly reducing the error of IESAm and IESAm$^+$ compared to IESA and IESA$^+$, respectively. On the other hand, the effect of moment matching decreases in M as the offered traffic is more finely divided into a larger number of request types. IESAm$^+$ is demonstrated to be the most accurate approximation for all values of λ shown.

6.6. *Varying the total number of server groups*

We consider a case with $L = 2$ layers, with $G_1 = 3n$ server groups in layer 1 and $G_2 = 2n$ servers in layer 2, for various values of n. We maintain $M = 500$, $k_{m,1} = k_{m,2} = n$ for all request types m, and $N_{\ell,g} = 10$ for all server groups (ℓ, g). The arrival rate is set so that the blocking probability is approximately 0.5% in all cases. The results are shown in Fig. 14.

The results demonstrate a rapid deterioration of FPA and FPAm as n (and thus $k_{m,1}$ and $k_{m,2}$) increases, due to the cascading effect in which approximation errors at lower-order overflows affect all higher-order overflow traffic. All four IESA approximations shown demonstrate a vast improvement in accuracy over FPA and FPAm, with an improvement of roughly seven orders of magnitude for $n = 50$. Moment matching is demonstrated to provide an additional small benefit, with IESAm$^+$ the most accurate of approximation for all values of n shown.

6.7. *Varying the number of servers per group*

We consider a case with $L = 2$ layers, $G_1 = 60$, and $G_2 = 40$ for all server groups (ℓ, g), $M = 500$, and $k_{m,1} = k_{m,2} = 20$ for all request types m. We set $N_{\ell,g} = N$ for various values of N for all server groups (ℓ, g). The arrival rate is set so that the blocking probability is approximately 0.5% in all cases. The results are shown in Fig. 15.

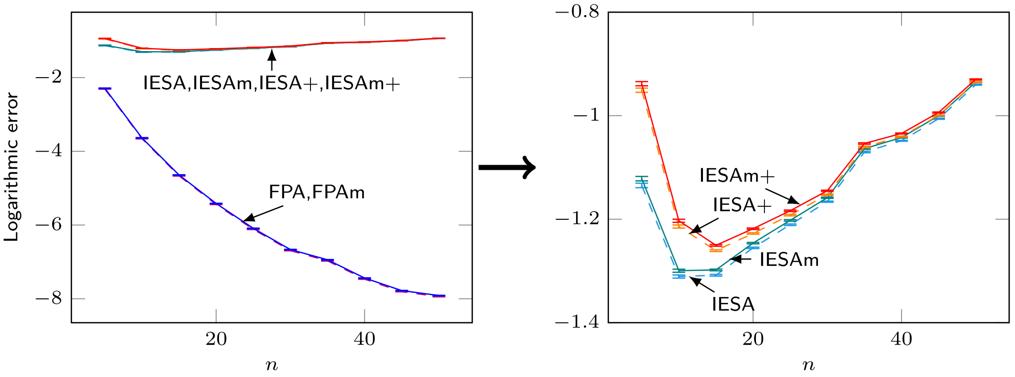

Fig. 14. Logarithmic errors for the scenarios described in Section 6.6. There are $3n$ server groups in layer 1 and $2n$ server groups in layer 2.

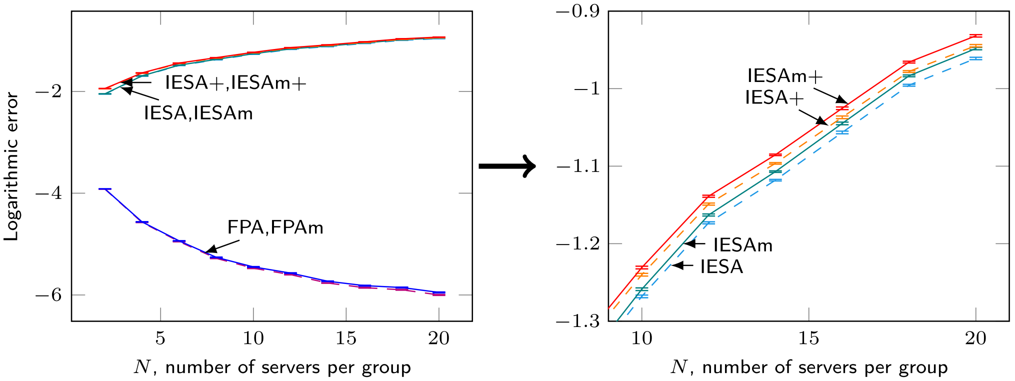

Fig. 15. Logarithmic errors for the scenarios described in Section 6.7.

The results demonstrate that while FPA and FPAm become less accurate as N increases, the IESA approximations become more accurate. Moment matching is demonstrated to provide an additional small benefit, with IESAm$^+$ the most accurate of approximation for all values of N shown. The effect of moment matching is largest when there is only one server per group.

6.8. Varying the arrival rate

In this subsection, we maintain $L = 2$, $G_1 = 60$, $G_2 = 40$, $N_{\ell,g} = 10$ for all server groups (ℓ, g), $M = 500$, and $k_{m,1} = k_{m,2} = 20$, while varying the arrival rate. Let λ denote the total arrival rate to the system, which is distributed evenly among all M request types. The results are shown in Fig. 16.

The results demonstrate a deterioration in accuracy of all the approximations as λ decreases. To explain this effect, we consider a simple M/M/k/k queue and note that for sufficiently loaded queues, the derivative of the Erlang B formula increases as λ decreases, as shown in Fig. 17. As a result, the estimate of the overflow probability of each server group becomes more sensitive to errors in estimating the offered load.

The IESA approximations are consistently more accurate than FPA and FPAm by several orders of magnitude. Moment matching is shown to have a small positive effect, slightly reducing the error of IESAm and IESAm$^+$ compared to IESA and IESA$^+$, respectively. IESAm$^+$ is demonstrated to be the most accurate approximation for all values of λ shown.

6.9. Unbalanced traffic

We consider a case with $L = 2$ layers, $G_1 = 60$, $G_2 = 40$, $N_{\ell,g} = 10$ for all server groups (ℓ, g), $M = 500$, and $k_{m,2} = 20$ for all request types m. On the other hand, $k_{m,1}$ is set to be proportional to λ_m with a mean of 20, subject to $1 \le k_{m,1} \le G_1$. The total arrival rate is set to $\sum_m \lambda_m = 960$, with $\lambda_m \propto m^{-0.6}$.

For each approximation, we post the distribution of the logarithmic error of each request type across twenty routing configurations. The results are shown in Fig. 18. For the sake of comparison, all bin widths are equal and are plotted on the same horizontal scale. The IESA approximations are demonstrated to reduce not only the approximation error for the mean blocking probability of the system, but also the spread of the error for individual request types.

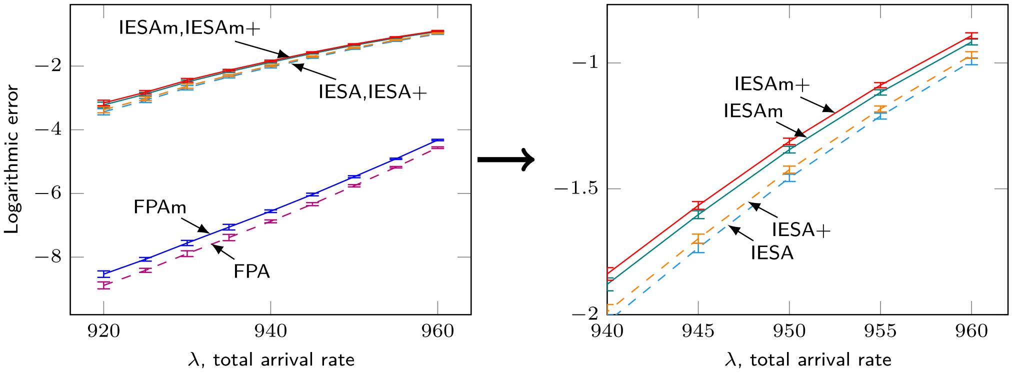

Fig. 16. Logarithmic errors for the scenarios described in Section 6.8.

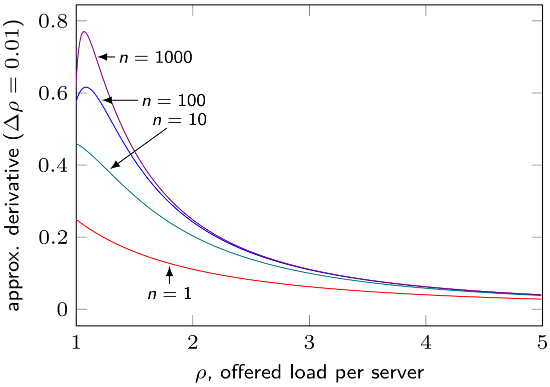

Fig. 17. Derivative of Erlang B formula with respect to the number of servers n, where $\rho = \frac{A}{n}$ is the offered load per server.

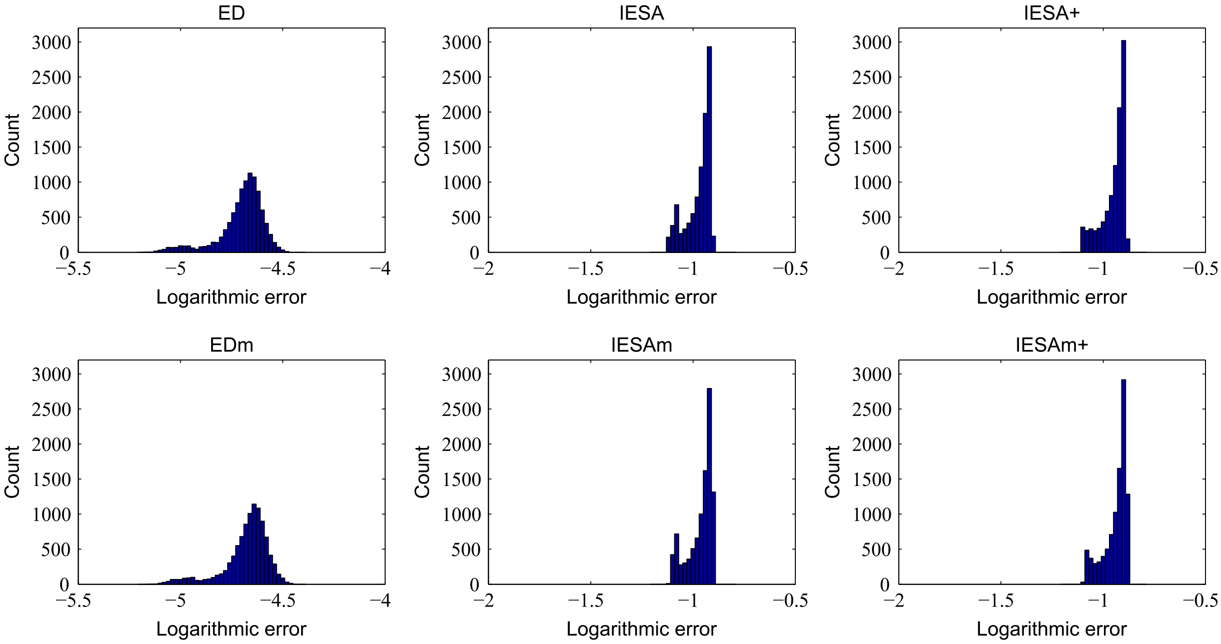

Fig. 18. Logarithmic errors for individual request types for an unbalanced system.

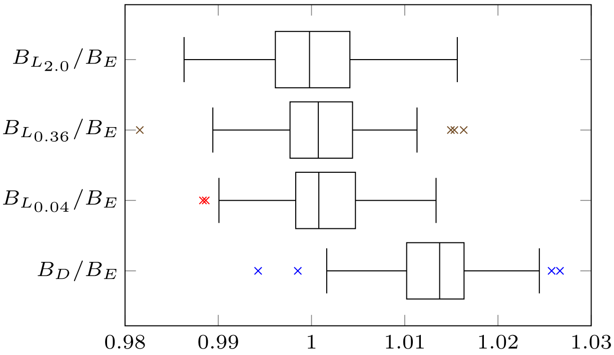

Fig. 19. Sensitivity of blocking probability to the service time distribution.

6.10. *Sensitivity to the service time distribution*

We consider the $k = 20$ case from Section 6.2, while changing the service time distribution of requests. Four distributions are considered, all with unit mean: exponential with a variance of 1.0, deterministic with a variance of 0.0, and lognormal with variances of 0.04, 0.36, and 2.0. The blocking probabilities for 100 randomly generated routing configurations are evaluated for each of the five distributions and denoted B_E, B_D, $B_{L_{0.04}}$, $B_{L_{0.36}}$, and $B_{L_{2.0}}$, respectively. The results, shown in Fig. 19, demonstrate less than 3% difference in B_D, $B_{L_{0.04}}$, $B_{L_{0.36}}$, and $B_{L_{2.0}}$ to B_E for all 100 configurations, suggesting that blocking probability in our overflow loss system model is not very sensitive to the service time distribution except through its mean. Whiskers show the maximum and minimum values within 1.5 times the interquartile range.

7. Concluding remarks

We have extended FPA and the IESA framework to a multi-layer loss system architecture with both hierarchical inter-layer overflow and non-hierarchical intra-layer overflow. We have proposed a new IESA approximation, IESA$^+$, which differs slightly from IESA on its handling of overflowing requests, based on the overflow history Δ and the congestion estimate Ω. We have strengthened FPA, IESA, and IESA$^+$ through the application of moment matching, generating FPAm, IESAm, and IESAm$^+$. Extensive numerical results demonstrate that IESA is consistently more accurate than FPA and FPAm, with improvements of several orders of magnitude in many cases. Furthermore, IESAm, IESA$^+$, and IESAm$^+$ provide an additional small but consistent improvement over IESA, with IESAm$^+$ providing the best results out of all the approximations considered. The IESA framework is shown to be most accurate when the number of layers is small, the number of accessible server groups per request is small, and the arrival rate is high.

Despite consistent improvement over conventional approximations such as FPA and FPAm, with several orders of magnitude in many cases, there are still cases where the IESA framework can be improved. Further work may be required to develop new IESA surrogates with increased accuracy and robustness, as well as moment matching techniques specifically tailored for IESA.

Finally, numerical results demonstrate near insensitivity of the blocking probability to the service time distribution except through its mean, allowing the IESA framework to be used in a wide range of overflow loss systems.

Acknowledgments

We thank Yu Peng for his contributions and comments in the early stages of this work.

This work was supported in part by grants from City University of Hong Kong (Project Nos. 9380044 and 7002860).

References

[1] N.M. van Dijk, Simple and insensitive bounds for a grading and an overflow model, Oper. Res. Lett. 6 (2) (1987) 73–76. http://dx.doi.org/10.1016/0167-6377(87)90033-2.

[2] K.S. Meier-Hellstern, The analysis of a queue arising in overflow models, IEEE Trans. Commun. 37 (4) (1989) 367–371. http://dx.doi.org/10.1109/26.20117.

[3] G. Falin, Asymptotic optimization of limited access queueing systems with losses, Perform. Eval. 26 (2) (1996) 77–93. http://dx.doi.org/10.1016/0166-5316(95)00018-6.

[4] N.M. van Dijk, E. van der Sluis, Call packing bound for overflow loss systems, Perform. Eval. 66 (1) (2009) 1–20. http://dx.doi.org/10.1016/j.peva.2008.06.003.

[5] D. McMillan, Traffic modelling and analysis for cellular mobile networks, in: Proc. 13th International Teletraffic Congress, ITC 13, 1991, pp. 627–632. URL http://www.itc-conference.org/fileadmin/ITCBibDatabase/1991/mcmillan911.pdf.

[6] K. Mitchell, K. Sohraby, An analysis of the effects of mobility on bandwidth allocation strategies in multi-class cellular wireless networks, in: Proc. IEEE INFOCOM 2001, vol. 2, 2001, pp. 1005–1011. http://dx.doi.org/10.1109/INFCOM.2001.916293.

[7] J. Wu, J. Guo, E.W.M. Wong, M. Zukerman, Approximation of blocking probabilities in mobile cellular networks with channel borrowing, in: Proc. IEEE 16th International Conference on High Performance Switching and Routing (HPSR), 2015, http://dx.doi.org/10.1109/HPSR.2015.7483075.

[8] V.O.K. Li, W. Liao, X. Qiu, E.W.M. Wong, Performance model of interactive video-on-demand systems, IEEE J. Sel. Areas Commun. 14 (6) (1996) 1099–1109. http://dx.doi.org/10.1109/49.508281.

[9] J. Guo, E.W.M. Wong, S. Chan, P. Taylor, M. Zukerman, K.-S. Tang, Performance analysis of resource selection schemes for a large scale video-on-demand system, IEEE Trans. Multimedia 10 (1) (2008) 153–159. http://dx.doi.org/10.1109/TMM.2007.911281.

[10] J.P. Muñoz-Gea, S. Traverso, E. Leonardi, Modeling and evaluation of multisource streaming strategies in P2P VoD systems, IEEE Trans. Consum. Electron. 58 (4) (2012) 1202–1210. http://dx.doi.org/10.1109/TCE.2012.6414986.

[11] R.C. Larson, Approximating the performance of urban emergency service systems, Oper. Res. 23 (5) (1975) 845–868. http://dx.doi.org/10.1287/opre.23.5.845.

[12] J.P. Jarvis, Approximating the equilibrium behavior of multi-server loss systems, Manage. Sci. 31 (2) (1985) 235–239. http://dx.doi.org/10.1287/mnsc.31.2.235.

[13] S. Budge, A. Ingolfsson, E. Erkut, Approximating vehicle dispatch probabilities for emergency service systems with location-specific service times and multiple units per location, Oper. Res. 57 (1) (2009) 251–255. http://dx.doi.org/10.1287/opre.1080.0591.

[14] M. Restrepo, S.G. Henderson, H. Topaloglu, Erlang loss models for the static deployment of ambulances, Health Care Manage. Sci. 12 (1) (2009) 67–79. http://dx.doi.org/10.1007/s10729-008-9077-4.

[15] J.B. Atkinson, I.N. Kovalenko, N. Kuznetsov, K.V. Mykhalevych, A hypercube queueing loss model with customer-dependent service rates, European J. Oper. Res. 191 (1) (2008) 223–239. http://dx.doi.org/10.1016/j.ejor.2007.08.014.

[16] N. Litvak, M. van Rijsbergen, R.J. Boucherie, M. van Houdenhoven, Managing the overflow of intensive care patients, European J. Oper. Res. 185 (3) (2008) 998–1010. http://dx.doi.org/10.1016/j.ejor.2006.08.021.

[17] M. Asaduzzaman, T.J. Chaussalet, N.J. Robertson, A loss network model with overflow for capacity planning of a neonatal unit, Ann. Oper. Res. 178 (1) (2010) 67–76. http://dx.doi.org/10.1007/s10479-009-0548-x.

[18] M. Asaduzzaman, T.J. Chaussalet, An overflow loss network model for capacity planning of a perinatal network, J. Roy. Statist. Soc. Ser. A 174 (2) (2011) 403–417. http://dx.doi.org/10.1111/j.1467-985X.2010.00669.x.

[19] G.F. O'Dell, An outline of the trunking aspect of automatic telephony, J. Inst. Electr. Eng. 65 (362) (1927) 185–215. http://dx.doi.org/10.1049/jiee-1.1927.0010.

[20] S.C. Graves, Flexibility principles, in: Building Intuition, in: International Series in Operations Research & Mngmt. Sci., vol. 115, Springer, 2008, pp. 33–49. http://dx.doi.org/10.1007/978-0-387-73699-0_3, (Chapter 3).

[21] J.P. Basset, P. Camus, 1000B Pentaconta crossbar switching system, ITT Electr. Commun. 38 (2) (1963) 196–212.

[22] R.J. Chapuis, 100 Years of Telephone Switching, Part 1: Manual and Electromechanical Switching (1878–1960's), second ed., IOS Press, 2003.

[23] B. Hennion, Feedback methods for calls allocation on the crossed traffic routing, in: Proc. 9th International Teletraffic Congress, ITC 9, 1979. URL http://www.itc-conference.org/fileadmin/ITCBibDatabase/1979/hennion79.pdf.

[24] F. le Gall, J. Bernussou, An analytical formulation for grade of service determination in telephone networks, IEEE Trans. Commun. 31 (3) (1983) 420–424. http://dx.doi.org/10.1109/TCOM.1983.1095813.

[25] U.R. Krieger, Analysis of a loss system with mutual overflow, in: Proc. ITC-Seminar, Peking, 1988.

[26] R.B. Cooper, S.S. Katz, Analysis of alternate routing networks with account taken of nonrandomness of overflow traffic, Memo. MM64-3122-2, Bell Telephone Laboratories (1964). URL http://www.cse.fau.edu/~bob/publications/Cooper_&_Katz_1964.pdf.

[27] F.P. Kelly, Blocking probabilities in large circuit-switched networks, Adv. Appl. Probab. 18 (2) (1986) 473–505. http://dx.doi.org/10.2307/1427309.

[28] R.I. Wilkinson, Theories for toll traffic engineering in the U.S.A., Bell Syst. Tech. J. 35 (2) (1956) 421–514. http://dx.doi.org/10.1002/j.1538-7305.1956.tb02388.x.

[29] A.A. Fredericks, Congestion in blocking systems – a simple approximation technique, Bell Syst. Tech. J. 59 (6) (1980) 805–827. http://dx.doi.org/10.1002/j.1538-7305.1980.tb03034.x.

[30] G.J. Franx, G. Koole, A. Pot, Approximating multi-skill blocking systems by hyperexponential decomposition, Perform. Eval. 63 (8) (2006) 799–824. http://dx.doi.org/10.1016/j.peva.2005.09.001.

[31] Q. Huang, K.-T. Ko, V.B. Iversen, An approximation method for multiservice loss performance in hierarchical networks, in: Proc. 20th International Teletraffic Congress, ITC 20, 2007, pp. 901–912. http://dx.doi.org/10.1007/978-3-540-72990-7_78.

[32] E.W.M. Wong, A. Zalesky, Z. Rosberg, M. Zukerman, A new method for approximating blocking probability in overflow loss networks, Comput. Netw. 51 (11) (2007) 2958–2975. http://dx.doi.org/10.1016/j.comnet.2006.12.007.

[33] E.W.M. Wong, J. Guo, B. Moran, M. Zukerman, Information exchange surrogates for approximation of blocking probabilities in overflow loss systems, in: Proc. 25th International Teletraffic Congress, ITC 25, 2013. http://dx.doi.org/10.1109/ITC.2013.6662932.

[34] Y.-C. Chan, J. Guo, E.W.M. Wong, M. Zukerman, Performance analysis for overflow loss systems of processor-sharing queues, in: Proc. IEEE INFOCOM '15, 2015, pp. 1409–1417. http://dx.doi.org/10.1109/INFOCOM.2015.7218518.

[35] A. Lotze, History and development of grading theory, in: Proc. 5th International Teletraffic Congress, ITC 5, 1967, pp. 148–161.

[36] P. Fitzpatrick, C.S. Lee, B. Warfield, Teletraffic performance of mobile radio networks with hierarchical cells and overflow, IEEE J. Sel. Areas Commun. 15 (8) (1997) 1549–1557. http://dx.doi.org/10.1109/49.634793.

[37] Q. Huang, K.-T. Ko, V.B. Iversen, Approximation of loss calculation for hierarchical networks with multiservice overflows, IEEE Trans. Commun. 56 (3) (2008) 466–473. http://dx.doi.org/10.1109/TCOMM.2008.060051.

[38] S. Kang, H. Yin, A hybrid CDN-P2P system for video-on-demand, in: Proc. 2nd International Conference on Future Networks, ICFN 2010, 2010, pp. 309–313. http://dx.doi.org/10.1109/ICFN.2010.83.

[39] K. Bandaru, K. Patiejunas, Under the hood: Facebook's cold storage system, May 2015. URL https://code.facebook.com/posts/1433093613662262/-under-the-hood-facebook-s-cold-storage-system-/.

[40] P. Chevalier, J.-C. Van den Schrieck, Optimizing the staffing and routing of small-size hierarchical call centers, Prod. Oper. Manage. 17 (3) (2008) 306–319. http://dx.doi.org/10.3401/poms.1080.0033.

[41] P. Chevalier, J.-C. Van den Schrieck, Approximating multiple class queueing models with loss models, ECORE Discussion Papers (2008) 33/1–23. URL http://ideas.repec.org/p/cor/louvco/2008021.html.

[42] Hong Kong Hospital Authority, Clusters, hospitals & institutions. URL http://www.ha.org.hk/visitor/ha_visitor_index.asp?Content_ID=10036.

[43] H.A. Longley, The efficiency of gradings, Post Off. Electr. Eng. J. 41 (1948) 45–49. 67–72.

[44] P.K. Das, D.G. Smith, Analysis of blocking probabilities in skipped gradings which are offered unbalanced traffic, IEE Proc. F: Commun. Radar Signal Process. 127 (6) (1980) 430–438. http://dx.doi.org/10.1049/ip-f-1.1980.0063.

[45] A.N. Avramidis, W. Chan, P. L'Ecuyer, Staffing multi-skill call centers via search methods and a performance approximation, IIE Trans. 41 (6) (2009) 483–497. http://dx.doi.org/10.1080/07408170802322986.

[46] G. Koole, J. Talim, Exponential approximation of multi-skill call centers architecture, in: Proc. 4th International Workshop on Queueing Networks with Finite Capacity, QNETs 2000, 2000, pp. 23/1–10.

[47] M. Schneps-Schneppe, J. Sedols, Multi-skill call center as a grading from "old" telephony, in: Proc. 9th International Conference on Next Generation Wired/Wireless Networking, NEW2AN 2009 and Second Conference on Smart Spaces, ruSMART 2009, 2009, pp. 154–167. http://dx.doi.org/10.1007/978-3-642-04190-7_15.

[48] E.W.M. Wong, B. Moran, A. Zalesky, Z. Rosberg, M. Zukerman, On the accuracy of the OPC approximation for a symmetric overflow loss model, Stoch. Models 29 (2) (2013) 149–189. http://dx.doi.org/10.1080/15326349.2013.783284.

[49] P. Chevalier, N. Tabordon, Overflow analysis and cross-trained servers, Int. J. Prod. Econ. 85 (1) (2003) 47–60. http://dx.doi.org/10.1016/S0925-5273(03)00085-9.

[50] A.A. Li, W. Whitt, Approximate blocking probabilities in loss models with independence and distribution assumptions relaxed, Perform. Eval. 80 (1) (2014) 82–101. http://dx.doi.org/10.1016/j.peva.2013.08.004.
[51] T.H. Burwell, J.P. Jarvis, M.A. McKnew, Modeling co-located servers and dispatch ties in the hypercube model, Comput. Oper. Res. 20 (2) (1993) 113–119. http://dx.doi.org/10.1016/0305-0548(93)90067-S.
[52] J.B. Atkinson, I.N. Kovalenko, N.Y. Kuznetsov, K.V. Mikhalevich, Heuristic methods for the analysis of a queuing system describing emergency medical service deployed along a highway, Cybernet. Systems Anal. 42 (3) (2006) 379–391. http://dx.doi.org/10.1007/s10559-006-0075-6.
[53] F.E. Browder, W.V. Petryshyn, The solution by iteration of nonlinear functional equations in Banach spaces, Bull. Amer. Math. Soc. 72 (3) (1966) 571–575. http://dx.doi.org/10.1090/S0002-9904-1966-11544-6.
[54] Y.C. Chan, Surrogate-based approximation of blocking probability in non-hierarchical overflow loss systems (unpublished thesis), 2016.
[55] E.L. Blair, C.E. Lawrence, A queueing network approach to health care planning with an application to burn care in New York State, Soc.-Econ. Plann. Sci. 15 (5) (1981) 207–216. http://dx.doi.org/10.1016/0038-0121(81)90041-0.
[56] J. Guo, Y. Wang, K.-S. Tang, S. Chan, E.W.M. Wong, P. Taylor, M. Zukerman, Evolutionary optimization of file assignment for a large-scale video-on-demand system, IEEE Trans. Knowl. Data Eng. 20 (6) (2008) 836–850. http://dx.doi.org/10.1109/TKDE.2007.190742.
[57] D.L. Jagerman, Methods in traffic calculations, AT&T Bell Labs Tech. J. 63 (7) (1984) 1283–1310. http://dx.doi.org/10.1002/j.1538-7305.1984.tb00037.x.
[58] E. Brockmeyer, H.L. Halstrøm, A. Jensen, The life and works of A. K. Erlang, no. 2 in Transactions of the Danish Academy of Technical Sciences, 1948.

Yin-Chi Chan received the B.Math. degree from the University of Waterloo, Waterloo, Ont., Canada, in 2010, and the M.Sc. degree from the City University of Hong Kong, Hong Kong in 2011. He is currently pursuing the Ph.D. degree in electronic engineering at City University of Hong Kong, Hong Kong.

His research interest is currently focused on approximative methods for the performance evaluation of stochastic loss systems.

Jun Guo received the B.E. degree in automatic control engineering from Shanghai University of Science and Technology, Shanghai, China, in 1992, and the M.E. degree in telecommunications engineering and the Ph.D. degree in electrical and electronic engineering from the University of Melbourne, Melbourne, Vic., Australia, in 2001 and 2006, respectively.

He was with the School of Computer Science and Engineering, University of New South Wales, Kensington, N.S.W., Australia, as a Senior Research Associate from 2006 to 2008 and on an Australian Postdoctoral Fellowship supported by the Australian Research Council from 2009 to 2011. Since 2012, he has been with the Department of Electronic Engineering, City University of Hong Kong, Kowloon Tong, Hong Kong, where he is currently a Senior Research Fellow. His research interests include green communications and networking, teletraffic theory and its applications in service sectors, and survivable network topology design.

Eric W.M. Wong 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 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 switching.

Moshe Zukerman received the B.Sc. degree in industrial engineering and management and the M.Sc. degree in operations research from the Technion-Israel Institute of Technology, Haifa, Israel, in 1976 and 1979, respectively, and the Ph.D. degree in engineering from the University of California, Los Angeles, CA, USA, in 1985. He was an Independent Consultant with the IRI Corporation and a Postdoctoral Fellow with the University of California, Los Angeles, CA, from 1985 to 1986. He was with the Telstra Research Laboratories (TRL), Melbourne, Vic., Australia, first as a Research Engineer from 1986 to 1988, and then as a Project Leader from 1988 to 1997. He also taught and supervised graduate students with Monash University, Melbourne, Vic., Australia, from 1990 to 2001. From 1997 to 2008, he was with the University of Melbourne, Melbourne, Vic., Australia. In 2008, he joined City University of Hong Kong, Kowloon Tong, Hong Kong, as a Chair Professor of Information Engineering and a Team Leader.

He has served on various Editorial Boards such as Computer Networks, the IEEE Communications Magazine, the IEEE Journal of Selected Areas in Communications, and the IEEE/ACM Transactions on Networking.