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), eewong@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

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